Localized Quantile Regression of Realized Volatility

by

© **Janaki Koralage**

A thesis submitted to the School of Gradate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University

January 2019

St. John's, Newfoundland and Labrador, Canada

Abstract

Volatility is a financial term that measures the dispersion of asset returns. Calculating and predicting volatility are not simple, but there are several well-known models for determining the volatility of assets. In recent years, researchers have been interested in developing statistical methods to model financial volatility, and new concepts have been applied to achieve better results. Quantile regression is another area gaining increased attention in the analysis of financial data. In this thesis, we propose a new quantile regression model for measuring the volatility of financial assets called the localized quantile regression model. As the name suggests, the proposed model is a local model rather than a global model. It takes care of possible structural changes and makes predictions of volatility more reliable. The initial step in this approach is to identify the longest interval of homogeneity. Identifying this interval of homogeneity involves a sequential testing procedure. After identifying intervals, we can apply quantile regression for each homogeneous time interval. The main advantage of this method is that it does not require any distributional assumptions. Simulation studies are carried out to investigate the performance of the proposed method. Results obtained from the simulation study show that the localized quantile regression model is appropriate for modeling the volatility of financial assets.

I dedicate this thesis to my mother, husband and my son for their endless love, support and encouragement throughout my life.

Lay summary

Assets are defined as property owned by a person or a company, such as money, stocks, bonds, real state, or investments. Usually, the prices of these assets change over time, with some, such as stock prices, changing very quickly. Many factors impact these price changes including the introduction of new company policies, changes in political and economic situations, etc. Every person or company that owns assets aims to earn a maximum profit by investing their assets in the market at the right time.

There are billions of investors in the stock market around the world. The main goal of these investors is to maximize profit while reducing the risk of losses. Stock prices can change instantaneously because investors adjust their decisions according to new reports based on past details about price changes. As a result, there can be large movements in the price of a given asset in a short period of time.

In this study, we investigate and build a model for variations in the price of a given asset. From period to period, there are unique, but sometimes different, patterns in these price variations. If we can successfully identify these time periods, we can build models separately for each period. Rather than modeling all data together, separately modeling may lead to more accurate predictions. This hypothesis is the main idea behind our research. Our approach is designated to handle high frequency financial data. However, we cannot access such data from any free sources. Therefore, we created data through simulation and applied the proposed method to the simulated data. The results show that the proposed method can be used to forecast price variations in a given asset.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Zhaozhi Fan, for his patience, enthusiasm, motivation, and immense knowledge. I could not have had a better supervisor for my studies. My co-advisor, Prof. J. C. Loredo-Osti, has always been there to listen and offer advice. Also special thanks for my husband, Prageeth Senadeera for his guidance and support in every way possible.I gratefully acknowledge the financial support in the form of graduate fellowships and teaching assistantships provided by Memorial University of Newfoundlands School of Graduate Studies, the Department of Mathematics and Statistics, and my supervisors. I wish to thank all the faculty in our department for their support. It would not have been possible for me to complete the requirements of the graduate program without their guidance. I would also like to thank my parents,brother and sister for their constant support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With recent developments in technology, we can now easily record all transactions in the financial market in greater detail, giving us massive amounts of financial data through the Internet and financial institutes. Financial data consist of information related to businesses such as banking, investments, assets, properties, liabilities, equity, and stocks. These data are essential to allow internal management to run businesses smoothly. Stockholders and organizations also need financial data to make decisions on investing in the market and to examine creditworthiness of a business. In addition, financial professionals need accurate and comprehensive financial information to achieve long-term goals and to make more reliable business decisions.

Methods and techniques used in statistical time series analysis are highly applicable for most financial data since they are available in time series formats. The primary objective of financial time series analysis is to determine the value of an asset over a given or desired period [22]. Financial time series analysis are highly empirical discipline, but we can use proven methods to make inferences [22]. Analyzing financial data is different from the analysis of other types of time series data, mainly because of the uncertainties associated with the financial data [22]. Rather than using price series, return series of assets are preferred in financial studies. Campbell et al. [3] presents two main reasons for this. The first reason is that assets series provide a scale-free and detailed summary for average investors. This information is helpful for deciding on investment opportunities. The second reason is that return series are easier to analyze than price series because return series have excellent statistical properties such as prices are bounded to be non-negative, but the log returns can range from positive infinity to negative infinity.

Return on Asset (ROA) is a good indicator of the efficiency of a company [22]. The ROA is defined by the net income of the company divided by the total assets; that is,

$$ROA = Net\ Income/Total\ Assets. \tag{1.1}$$

This ratio gives us the percentage of the profit a company earns compared to its resources. ROA is sometimes referred to "Return on Investment". There are several types of asset returns such as the one-period simple return, multi-period simple return, and continuous compounded return (log return) [22]. If we take P_t to be the price of an asset at time t, we can define the one-period simple return at time t (R_t) as, $1 + R_t = \frac{P_t}{P_{t-1}}$, or

$$(1 + R_t)P_{t-1} = P_t. \tag{1.2}$$

The continuously compounded return denoted by r_t is the natural logarithm of the simple gross return. This return is also called the log return, and is shown by:

$$r_t = \ln(1 + R_t) = \ln(P_t) - \ln(P_{t-1}). \tag{1.3}$$

Volatility is another important measurement in finance and related fields. Returns are used to calculate the volatility of an asset. Volatility (σ) is defined as a statistical measure of the degree of variation of a trading price series over time. If there are m returns, $\hat{\sigma}$, an empirical estimator of σ, can be calculated using the standard deviation of logarithmic returns as

$$\hat{\sigma}^2 = \frac{1}{m-1}\sum_{t=1}^{m}(r_t - \bar{r_t})^2, \tag{1.4}$$

where $r_t = ln\frac{P_t}{P_{t-1}}$ (r_t is the continuously compounded return on asset during the day t), $\bar{r}_t$ is the mean of all m returns in the sample, and P_t is the market value of the asset on the day t. Usually, the standard deviation of past returns is a natural estimator of σ.

We can identify the realized volatility and the implied volatility as the two main types of volatilities. Realized volatility (also called historical volatility) is calculated using the historical market data, therefore it is known. Unlike realized volatility implied volatility is calculated using option prices and is used to predict the future volatility of a stock. In real market data, volatility can take values from zero to one hundred percent (0 to 100%), but theoretically, it can spread from zero to positive infinity. Volatility is zero when the price of the asset is constant over time. Volatilities

above 100% are rare.

The main difficulty in calculating volatility is that we cannot directly observe it. To estimate daily volatility, we need frequent data such as returns in every 10-minutes. However, we cannot guarantee the accuracy of such estimates. For instance, stock volatility consists of intraday volatility and overnight volatility where the latter denotes variation between trading days. High-frequency intraday returns contain limited information about overnight volatility. Due to these restrictions it is hard to make inferences about the future performance of volatility.

One of the main purposes of analyzing financial data is to minimize the associated risk [7], [17]. Usually, the higher the volatility, the higher the risk [4]. Volatility forecasting is widely used in risk management, derivative pricing, hedging market making, market timing, and portfolio selection [4], [7]. Several studies have been conducted over the past few years to examine the characteristics of volatility. There are also many empirical findings on volatility. These findings are consistent so that they are called stylized facts about volatility [7].

Engle and Patton [7] have mainly focused on a few common characteristics (also called stylized facts) of the volatility. The first property of the volatility series is that it exhibits persistence [7], [17]. Volatility is not a constant, and tends to cluster in time. Observing a large (or small) return today (regardless of the sign) is a good precursor of large (or small) returns in the upcoming days. Mandelbrot [16] and Fama [8] describe that large changes in the price of an asset are followed by other large changes, and small changes are followed by other small changes of price of an asset. These changes can lead volatility to cluster over time. Changes in volatility typically have very long-lasting impacts on its subsequent evolution. A number of studies have reported this behavior, such as Baillie et al. [1], Chou [5], and Schwert [21].

Volatility has a long memory. Volatility is not a constant for long periods because it fluctuates often. These fluctuations will finally converge to a normal level of volatility. This characteristic is called the mean reverting. Innovations may have an asymmetric impact on volatility; that is, most volatility models work under the assumption that positive and negative innovations (changes in returns) symmetrically affects the conditional volatility. For instance, the GARCH (1,1) model assumes that variance is affected by the square of the lagged innovations. Therefore, the GARCH

(1,1) model does not consider the sign of the innovations. However, positive and negative shocks do not have the same impact on volatility. If the volatility is high, the demand for a stock is low. In other words, high volatility reduces the value of the stock. This property is also referred to as the leverage effect. The prices of financial assets are correlated with other variables and deterministic events may also have an impact on volatility. This means that volatility is not independent of other exogenous variables.

There are two categories of volatility models. In the first category, the conditional variance is modeled directly as a function of observables [7], [17]. ARCH and GARCH types of models fall into this category. In the second category, models of volatility usually requires restricted model specifications. To this end, a stochastic equation is used to describe σ_t^2. Therefore, the second category is called latent volatility or stochastic volatility. A few of the commonly used models are listed below:

- EWMA -Exponentially Weighted Moving Average
- GARCH- Generalized Autoregressive Conditional Heteroskedasticity
- EGARCH- Exponential GARCH
- Regime-Switching GARCH
- FIGARCH- Fractionally Integrated GARCH
- SWARCH-Switching ARCH.

Chen et al. [4] have proposed a localized modeling approach for the realized volatility by introducing a time-varying (local) structure of volatility. A unique feature of this approach is that it uses an adaptive statistical technique to determine these time-varying structures. The model is based on one assumption that at each time point there exists a past-time interval (interval of homogeneity) where volatility can be described by a local autoregressive (LAR) model. This approach is local rather than global. Therefore, the length of the past time interval and parameters are time-varying and are different from one such period to another.

Quantile regression is another approach to model and predict the volatility and does not require any distributional assumptions. Koenker and Bassett [14] introduced the quantile regression in 1978. It has since become a widely-used technique in many

research areas such as medical reference charts, survival analysis, financial economics, and environment modeling [23].

Linear regression summarizes the average relationship between a set of regressors (Xs) and an outcome variable (Y) based on the conditional mean, denoted by $E(Y|X)$. However, it does not provide a full picture of this relationship in a sense that one may want to look at the relationship at different points in the conditional distribution of Y given a set of regressor X=x. Quantile regression plays an essential role in these types of situations. There are other advantages of selecting quantile regression over linear regression such as quantile regression is robust to response outliers and it is invariant to monotonic transformations. Quantile regression has recently been extended to financial data analysis. Huang [12] proposed a new method using quantile regression. Instead of using one pair of quantiles, he used a uniformly-spaced series of quantiles to describe volatility.

In this study, we propose a localized quantile regression approach. The proposed approach sequentially identifies homogeneous intervals and then applies a quantile regression model to each homogeneous interval. The quantile regression model hence has time-dependent coefficients. The quantile regression model does not require distributional assumptions. Direct interpretation of the results at selected quantiles might be of more interest to practitioners in the area of finance. The simulation study presented in Chapter 4 shows that the localized quantile regression model fits the realized volatility more closely and is also more predictive, as compared to its global counterpart.

Chapter 2

Quantile Regression

Quantile is a data summarization measurement used in statistics and is also known as the percentile. A median (50^{th} percentile) is an example of a quantile which partitions the observations into two equal parts. We denote the quantile value by q_τ, where τ lies between zero and one ($0 < \tau < 1$). The computation and interpretation of quantiles are straightforward. For example, when τ is equal to 0.5, it gives the median of the data, which tells us that half of the data is above the median and half is below the median. Another simple example is the growth chart of babies. For example, if a four-month-old baby is in the 60^{th} percentile for weight; that means 60 percent of four-month-old babies weigh the same as or less than that baby, and 40 percent weigh more.

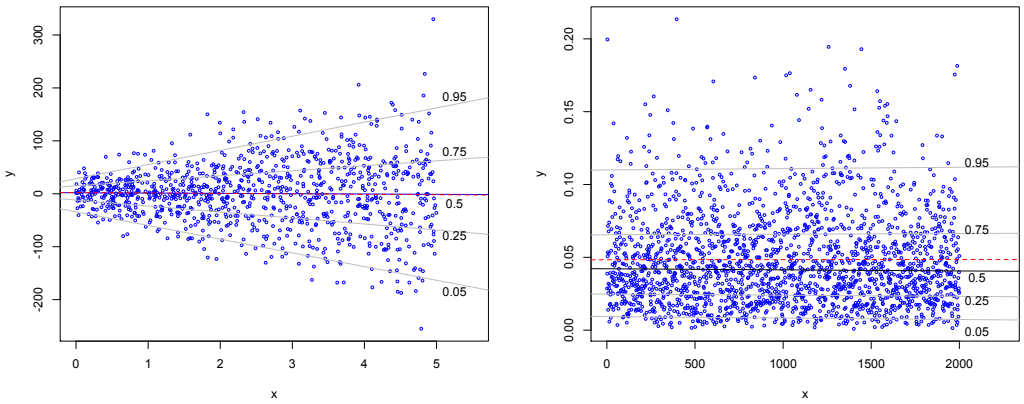
Koenker and Bassett [14] introduced the quantile regression as a method of statistical modeling, and since then it has become widely used in many areas. Unlike linear regression, quantile regression is more robust to outliers, and does not require to assume a constant variance for the response. Moreover, a fundamental assumption of the traditional linear regression approach is to assign a distribution for response variable, but quantile regression does not assume a parametric distribution. Quantile regression has the capability of describing the entire conditional distribution of response by estimating models for the conditional distribution of response for different quantile levels.

The traditional linear regression method provides only a summary of the relationship between the mean of the response variable (Y) and the set of regressors (Xs). It estimates a model for the conditional mean function. Linear regression parameters are determined using the least squares estimation technique, which minimizes the sums

of squares of the residuals. This conditional mean curve ($E(Y|X)$) does not have any potential to describe the entire conditional distribution. For example, classical least squares regression cannot be used to adequately characterize the relationship between response and covariates when data come from a skewed distribution or when heteroscedasticity exists in the data.

However, for different quantiles, we can draw different quantile regression lines, which are unique for each quantile value. We can model the entire conditional distribution of the response variable by drawing a series of regression lines for different values of τ. In Figure 2.1, the graph (a) shows the quantile regression lines from a data set where the mean and the variance of a response variable increase with the value of a covariate, and the graph (b) shows a data set which is asymmetric around the mean. Both graphs provide examples in which linear regression does not capture relevant information contained in the data such as heteroscedasticity. Both figures show the fitted quantile regression lines for the quantile levels 0.05, 0.25, 0.5, 0.75, and 0.95. The red line is the least squares regression line. As we can see from the graphs, the least square regression line does not capture the conditional variance. When heteroscedasticity exists in the data, we stabilize the variance using relevant monotone transformation. This will lead us to ignore valuable information contained in the data. Moreover, quantile regression does not require the assumption of a constant variance. Therefore, stabilizing the variance is not necessary for building quantile regression models and quantile regression is invariant to monotonic transformations such that $Q_\tau(h(y|x)) = h(Q_\tau(y|x))$, where $h(.)$ is a monotone increasing function and $Q_\tau(y|x)$ denote the τ^{th} conditional quantile of y given x [15].

Quantile regression has been used in medical studies, financial and economics research, and environment modeling. It has many applications in medical sciences such as in growth and reference charts of the height and weight of children. It has the potential to identify unusual subjects lying in the upper tail or the lower tail. Further investigation can be done to identify the factors related to the cases and hence make medical diagnoses more reliable. Also, quantile regression is used in economics to determine wages, discrimination effects, and trends in income [15].



(a) (b)

Figure 2.1: Regression models for different quantile levels

2.1 Quantiles and Quantile Functions

Let $Q_\tau(Y|X)$ denote the τ^{th} conditional quantile of Y given X. The quantile level τ is the probability $Pr[Y \leqslant Q_\tau(Y|X)|X]$, which gives the value of Y, below which the proportion of the conditional response population is τ. Any real-valued random variable, Y, may be characterized by its distribution function,

$$F(y) = P(Y \leqslant y), \tag{2.1}$$

whereas, for any $0 < \tau < 1$,

$$Q_\tau(Y) = F^{-1}(\tau) = inf\{y : F(y) \geqslant \tau\}, \tag{2.2}$$

is called the τ^{th} quantile of Y. To obtain sample quantile, we replacing F by the empirical distribution function

$$F_n(y) = n^{-1}\Sigma_{i=1}^{n} I(Y_i \leqslant x). \tag{2.3}$$

Then the τ^{th} sample quantile given as

$$\hat{Q}_\tau(Y) = F_n^{-1}(\tau) = inf\{y : F_n(y) \geqslant \tau\}. \tag{2.4}$$

Let Y be the response variable and X be an $n \times p$ regressors matrix, consisting of the $p \times 1$ vector of covariates x_i for the i^{th} observation for $i = 1, 2, \ldots, n$. Then, the statistical regression model for the linear relationship between the response and covariates can be given as $Y_i = x_i^T \beta + \epsilon_i$, where ϵ_i is the error term and β is a $p \times 1$ vector of unknown parameters. We assume that the ϵ_i are independent and identically normally distributed with mean 0 and variance σ^2; that is, in notation, $N(0, \sigma^2)$. Using the above information, we can give the average response as,

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{2.5}$$

where $i = 1, \ldots, n$. The model parameters, $\beta_1, \beta_2, \ldots, \beta_p$, can be estimated by minimizing the summation of the squared model prediction errors ϵ_i; that is $\sum_i \epsilon_i^2$. We define such an estimate of the vector of parameters β by $\hat{\beta}$, which is defined as follows

$$\hat{\beta} = argmin_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2, \tag{2.6}$$

or in the matrix form

$$\hat{\beta} = argmin_{\beta \in \mathbb{R}} \sum \left(y_i - x_i^T \beta \right)^2. \tag{2.7}$$

The linear conditional quantile model can be formulated as $Q_y(\tau | x_i) = x_i^T \beta(\tau)$, which gives quantile regression model as

$$y_i = x_i^T \beta(\tau) + \epsilon_i, \tag{2.8}$$

where $\beta(\tau)$ is $p \times 1$ unknown quantile regression parameters of interest at the τ^{th} quantile, ϵ_i is the error term with the density (say, $f_p(\cdot)$) and $i = 1, \ldots, n$. In the density $f_p(\cdot)$, τ^{th} quantile of ϵ_i is zero; that is, $\int_{-\infty}^{0} f_p(\epsilon_i) d\epsilon_i = \tau$ [20]. Then, the linear conditional quantile function is

$$Q_i(\tau | x_i) = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \ldots + \beta_p(\tau) x_{ip}, \tag{2.9}$$

where $i = 1, \ldots, n$.

The estimation technique in the quantile regression is different from the least squares technique. It estimates quantile regression model parameters by minimizing

a summation that gives asymmetric penalties $(1-\tau)|\epsilon_i|$ for overprediction and $\tau|\epsilon_i|$ for underprediction. To estimate the quantile regression model given in the equation (2.9), linear programming methods are used [20], [13]. To be specific, the β_j can be estimated by solving the minimization problem

$$\hat{\beta}_\tau = argmin_{\beta_1(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau\Big(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau)\Big), \tag{2.10}$$

or, in the matrix form,

$$\hat{\beta}_\tau = argmin_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_\tau\Big(y_i - x_i^T\beta_\tau\Big), \tag{2.11}$$

where the function ρ_τ is referred to as the check loss function. Its shape looks like a check mark. For any $0 < \tau < 1$, we define the piecewise linear "check loss function", $\rho_\tau(u) = u(\tau - I(u < 0))$, where $I\{\cdot\}$ is the usual 0-1 valued indicator function [13]. The check loss function is illustrated in Figure 2.2. It should be note that the quantile regression estimators given in equation (2.11) are asymptotically normally distributed [14].

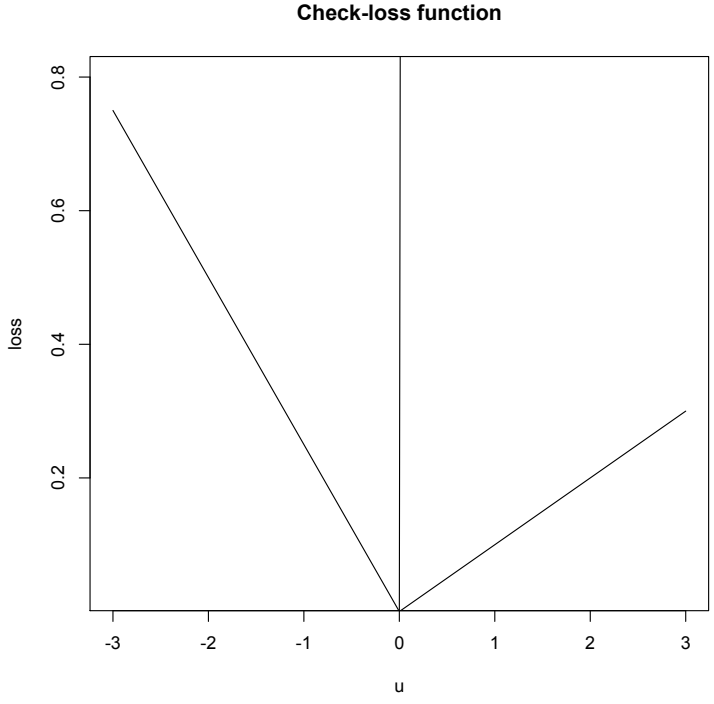


Figure 2.2: Check loss function

2.1.1 Equivariance and Transformation

Equivariance properties are often treated as an important aid in interpreting statistical results. Let the τ^{th} regression quantile based on observations (y, X) be denoted by $\hat{\beta}_\tau(y, X)$. We collect some basic properties in the following result: Proposition 2.2.1 (Koenker and Bassett, 1978) [14]. Let A be any $p \times p$ nonsingular matrix, $\gamma \in \mathbb{R}^p$, and $a > 0$. Then, for any $\tau \in [0, 1]$,

1. Scale equivalence: For any $a > 0$,
   - $\hat{\beta}_\tau(-ay, X) = -a\hat{\beta}_{1-\tau}(y, X)$
   - $\hat{\beta}_\tau(ay, X) = a\hat{\beta}_\tau(y, X)$
2. Regression shift: For any $\gamma \in \mathbb{R}^p$,
   - $\hat{\beta}_\tau(y + X\gamma, X) = \hat{\beta}_\tau(y, X) + \gamma$
3. Reparameterization of design: For any $|A| \neq 0$,
   - $\hat{\beta}_\tau(y, XA) = A^{-1}\hat{\beta}_\tau(y, X)$

2.2 Quantile Regression Using Asymmetric Laplace Distribution

The asymmetric Laplace distribution (ALD) is a generalization of the Laplace distribution. We say a random variable Y follows an asymmetric Laplace Distribution, if its probability density function (pdf) is

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\big\{ -\rho_\tau\big(\frac{y-\mu}{\sigma}\big)\big\}, \tag{2.12}$$

where μ is the location parameter, σ is the scale parameter ($\sigma > 0$), and $\tau \in (0, 1)$ is the skewness parameter. Here, ρ_τ is the check loss function, which is illustrated in Figure 2.2 for a specific τ value. We use $ALD(\mu, \sigma, \tau)$ to denote the distribution of asymmetric Laplace distribution with relevant parameters. If $W = \rho_\tau(\frac{Y-\mu}{\sigma})$, it is easy to see that W follows an exponential distribution with mean σ [20]. Figure 2.3 shows the probability density function of some ALD's.

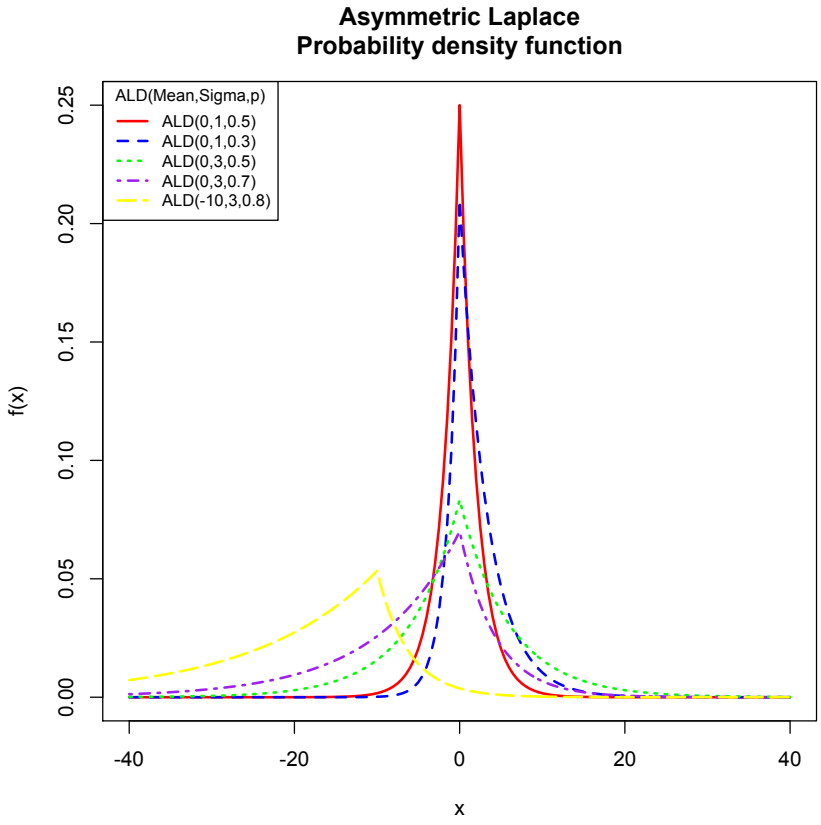


Figure 2.3: Probability density function of asymmetric Laplace distribution, Mean is the location parameter, Sigma is the scale parameter, and p is the skewness parameter

Suppose that $y_i \sim ALD(X_i^T\beta_\tau, \sigma, \tau)$, $i = 1, \ldots, n$. Likelihood function based on n independent observations from (2.12) is given by,

$$L(\beta, \sigma|y) = \frac{\tau^n(1-\tau)^n}{\sigma^n} exp\{-\sum_{i=1}^{n} \rho_\tau\big(\frac{y_i - X_i^T\beta_\tau}{\sigma}\big)\}. \tag{2.13}$$

If we take σ as a nuisance parameter, then the maximization of the likelihood in equation (2.13) with respect to the β_τ is equivalent to the minimization of the objective function in equation (2.11). This links the relationship between ALD and the inference for quantile regression estimation. Therefore, the relationship between ALD and the check loss function can be used in quantile regression studies [9].

Chapter 3

Models for Volatility

Appropriately modeling data is never an easy task, and finding a good model is always a challenge for statisticians. Modeling and predicting volatility are extremely difficult due to the complex behavior of data. Many studies have been conducted to model and predict the volatility of assets. Researchers have also proposed several models to capture the unique characteristics of volatility.

In this chapter, we present frequently used volatility models and propose a new model that uses quantile regression for volatility. A natural estimator of the volatility of financial assets is the standard deviation (σ) of past returns. There are a number of models that use the past values of returns to estimate volatility. A commonly used one is the historical average; this calculates $\hat{\sigma}_{t-1}$ and assumes it is equal to $\hat{\sigma}_t$. The drawback of this method is that its success depends on the sample size. Small sample sizes can lead to a large sampling error, while if we select a large sample, it may contain data which may not be relevant to the current market conditions. Exponentially Weighted Moving Average method was introduce to select a reasonable sample size to calculate volatility (σ) to reflect current market conditions. By assigning larger weights to more recent data, this model removes unnecessary information coming from more remote data.

3.1 Volatility Modeling and Prediction by ARCH and GARCH

In this section, we discuss statistical models that are commonly used in financial time series. These models can capture stochastic financial volatility, mean reversion,

and excess kurtosis. One of the experimentally proven characteristics of financial time series is that high-frequency time series tend to have fatter tails than Gaussian distribution [22]. The term "fatter tail" is also referred to as "excess kurtosis" [17]. The kurtosis of the distribution is a good indicator of whether unusual events have occurred outside the normal range. If the distribution has excess kurtosis, it tells us that there are many instances of outliers. These extreme values lead the distribution to have a fat tail on the bell-shaped distribution curve. Generally, if the kurtosis coefficient is more significant than the coefficient of a normal distribution, which is around 3, this indicates there is a possibility of obtaining more extreme outcomes than are usually found in normal distribution outcomes.

An Autoregressive Moving Average (ARMA) model has two major components. One is from an AR model with order p and the other is from a MA model with order q, denoted as ARMA(p,q). Let y_t be the stationary time series that is a realization of a stochastic process [17]. Mathematically, the ARMA(p,q) model can be expressed as [22]

$$y_t = \xi + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t - \sum_{j=1}^{q} \theta_i \epsilon_{t-j}, \tag{3.1}$$

where $\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2} \ldots$, are independent random errors, ξ is a constant, and ϕ_i and θ_i are autoregressive and moving average model parameters, respectively. When the series is not stationary, the ARIMA model can be used [17]. If y_t is a non-stationary time series, ARIMA(p,d,q) can be defined as,

$$\Phi(B)(1-B)^d y_t = \mu + \Theta(B)\epsilon_t, \tag{3.2}$$

where d is the order of differentation, Φ and Θ are autoregressive and moving average model parameter vectors, respectively, μ is mean of the y_t series, and B is the backshift operator, also referred to as the lag operator. As an example, if $BY_t = Y_{t-1}$, then B^d is the backshift d times ($B^d Y_t = Y_{t-d}$). When d takes the value zero, the ARIMA model becomes an ARMA model [17].

Financial time series have fatter tails. Hence, ARIMA models are not suitable for modeling this kind of time series, because they do not have the ability of capturing stochastic non-constant volatility [17]. Engle [6] proposed a solution to this problem by suggesting a new model called an ARCH (Autoregressive Conditional Heteroskedastic) model. As the name implies, this model assumes heteroskedastic variance. The ARCH model can capture the nonconstant volatility [6]. The ARCH model was very popular

when it was introduced by Engle [6]. It is the first model to assume that the volatility is not constant.

ARCH models well describe the fundamental properties of volatility such as volatility clustering, mean reversion, fat tails, and some other stochastic properties of volatility. Assume that y_t follows the standard stationary autoregressive model with order p; that is,

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-1} + \epsilon_t, \tag{3.3}$$

where ϵ_t is the residual with mean zero. We assume $|\phi_i| < 1$ for $i = 1, 2, \ldots, p$. The mathematical model of ARCH(p) [6] can be rewritten as,

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \alpha_1 a_{t-1}^2 + \ldots + \alpha_m a_{t-p}^2, \tag{3.4}$$

where a_t is the asset return at time t, σ_t is the volatility at time t, ϵ_t identically and independently distributed as $N(0, 1)$, $\omega > 0$, $\alpha_i \geqslant 0$ for $i = 1, 2, \ldots, m$.

Building an ARCH model to accurately estimate the volatility can be sometimes a challenge because there are several restrictions that have to be imposed. To deal with this, Bollerslev [2] proposed a generalized ARCH model (GARCH). Since then, GARCH has become frequently used in volatility forecasting [12]. Let r_t be the continuously compounded log return series and a_t the innovation of the series at time t defined as $a_t = r_t - \mu$. [17] The functional form of the GARCH(m,s) model is

$$a_t = \sigma_t \epsilon_t,$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{3.5}$$

where $\epsilon_t \sim N(0, 1)$ iid, $\omega > 0$, α_i and β_j are ARCH and GARCH parameters, respectively, $\alpha_i \geqslant 0$, $\beta_j \geqslant 0$, and $\sum_{i=1}^{max(m,s)} (\alpha_i + \beta_i) < 1$.

GARCH, IGARCH, TGARCH, EGARCH, and GARCH-M are all transformed models based on Engle's basic ARCH model. Some models are introduced to capture specific characteristics of volatility. As an example, the Threshold GARCH (TGARCH) model was proposed to capture the negative movements of volatility, which are usually larger than the positive movements. The TGARCH model, introduced by Glosten et al. [10] and Zakoian [24], is also frequently used in volatility

modeling. The main concern in this model is that it captures the movements of negative shocks because effects coming from negative movements are larger than from positive shocks [22]. Nelson [19] suggested a model called the exponential GARCH (EGARCH), which allows for unequal changes in volatility. Again, suppose a_t denotes the innovation of the asset return at time t, then we can write the EGARCH (m,s) model as:

$$a_t = \sigma_t \epsilon_t$$

$$\log(\sigma_t^2) = \omega + \sum_{i=1}^{s} \alpha_i \frac{|a_{t-i}| + \theta_i a_{t-i}}{\sigma_{t-i}} + \sum_{j=1}^{m} \beta_j \sigma_{t-j}^2. \tag{3.6}$$

The GARCH model gives highly accurate results in the long horizon when compared with complex models like MRS-GARCH (Markov Regime-Switching GARCH). But in a short time horizon, MRS-GARCH performs better than the GARCH model [18]. Regime Switching GARCH (RS-GARCH) and Fractionally Integrated GARCH (FIGARCH) are also commonly-used models introduced by Gray [11] and Baillie et al. [1], respectively.

3.2 Localized Realized Volatility Modeling

Another interesting approach to modeling volatility is introduced by Chen et al. [4]. The main purpose of their study was the investigation of the dual view on volatility. The proposed method is very flexible because it can explain short memory processes with breaks as well as long memory processes. Long memory processes can be diagnosed using the sample autocorrelation function (ACF). In other words, if the shape of the ACF is more hyperbolically decaying, then it can indicate long memory property of volatility.

Since the approach of the Chen et al. [4] is local rather than global, it is called localized realized volatility modeling. The idea is as follows. First, an algorithm is developed to identify homogeneous time intervals. Then, for each time interval, they proposed a local autoregressive (LAR) model because the main assumption behind this model is that at each time point there is a past time interval, where the data can be explained by the LAR model. Therefore, at each point in time, where there is a structural break, a new set of parameters is estimated. To this end, the maximum likelihood method is used to estimate these local parameters. The selection of the interval of homogeneity is a sequential testing procedure. First, it starts with a small

interval, where the local approximation is compatible and the estimated autoregressive parameters are approximately constant. Then, the algorithm iteratively expands this span and tests for time homogeneity. The procedure is repeated until a structural break is found or data are exhausted. Finally, the LAR model given in equation (3.7) is fitted.

Let the local autoregressive parameter set at time t be $\theta_t = (\theta_{0t}, \theta_{1t}, \ldots, \theta_{pt})^T$. These parameters are time-varying. The LAR model is

$$\log\ RV_t = \theta_{0t} + \sum_{i=1}^{p} \theta_{it} \log\ RV_{t-i} + \epsilon_t, \tag{3.7}$$

where ϵ_t follows normal distribution with mean zero and variance σ_t^2, and RV_t is the realized volatility at time point t. In the model (3.7) the log of RV was modeled because the realized volatility distribution is strongly skewed and has a fat tail, but the log of the realized volatility approximately follows a normal distribution [4]. In this model, all the parameters and the length of the past time intervals are different from interval to interval.

3.2.1 Parameter Estimation

To estimate the parameters in the LAR model (equation (3.7)) the maximum likelihood (ML) techniques can be used [4]. For a given homogeneous time interval (say I_τ for the time point τ), the ML estimator of θ_τ is

$$\tilde{\theta}_\tau = \underset{\theta \in \Theta}{\operatorname{argmax}}\ L(log\ RV; I_\tau, \theta),$$

$$\tilde{\theta}_\tau = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ -\frac{l_\tau - p}{2} \log\ 2\pi - (l_\tau - p) \log\ \sigma - \frac{1}{2\sigma^2} \sum_{t=\tau - l_\tau + p}^{\tau - 1} \left(\log\ RV_t - \theta_0 - \sum_{i=1}^{p} \theta_i \log\ RV_{t-i} \right)^2 \right\}, \tag{3.8}$$

where Θ is the parameter space and I_τ is the interval of homogeneity at time point τ, i.e. $I_\tau = [\tau - l_\tau, \tau)$ with $p + 2 \leqslant l_\tau < \tau$, p is the number of parameters [4].

3.3 Localized Quantile Regression of Realized Volatility

The extreme quantiles of volatility are of great importance in risk management. Motivated by this, we propose a localized quantile regression approach to volatility modeling and forecasting. The proposed approach sequentially identifies homogeneous intervals, and then applies a quantile regression model to each homogeneous interval. Hence, the quantile regression model has time-dependent coefficients. The proposed quantile regression model is

$$log\ RV_t = \beta_{0\tau} + \sum_{i=1}^{p} \beta_{i\tau} log\ RV_{t-i} + \epsilon_t, \tag{3.9}$$

where RV_t is the realized volatility calculate at time t, $\beta_{0\tau}$ and $\beta_{i\tau}$ are quantile regression parameters , and ϵ_τ is the error term where the τ^{th} quantile of ϵ_τ is zero.

3.3.1 Parameter Estimation

Linear programming methods are usually required to estimate quantile regression model parameters. One of the advantages of using linear programming methods is that these methods are computationally efficient. The parameters of the equation (3.9) can be estimated by solving the objective function given as,

$$\hat{\beta}_\tau = \min_{\beta_1(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau \left(log\ RV_i - \beta_0(\tau) - \sum_{j=1}^{p} log\ RV_{i-j} \beta_j(\tau) \right), \tag{3.10}$$

where the function ρ_τ is the check loss function. Under some regularity conditions, the estimators of quantile regression parameters are asymptotically normally distributed [13].

Various linear programming methods have been proposed to estimate quantile regression parameters. These methods include, for instance, the simplex method, interior point method, interior point method with preprocessing, and smoothing method. All these methods are different from each other. As an example, the simplex method starts from a random vertex, and then moves along the boundaries of a circumscribed polygon until the optimum solution is reached. The simplex method is ideal for moderate-sized data sets. The interior point method starts from an internal point

of the circumscribed polygon and looks for the optimum solution within the polygon without touching boundaries. The interior point method can be used for large data sets because it is computationally efficient.

We can easily estimate quantile regression parameters using the packages built in `R` and `SAS`. In our study, we used the package `Quantreg`, which supports the quantile regression in `R`.

Chapter 4

Simulation Study

We conducted simulation studies to investigate how well the quantile regression captures the volatility of a financial data set. In this chapter, we present the simulation setup and detailed the results. There are two main sections in the simulation study. In the first part, we applied the quantile regression approach to a generated data set. In the second part, we carried out another simulation study to find the parameters of the interval of homogeneity. We generated data from the localized autoregressive LAR(1) model equation (3.9) and applied the proposed localized quantile regression model as an illustration.

4.1 Simulation Setup

We first introduce the simulation setup.

- Generate data from LAR(1) processes with $\theta_t = \theta^* = (\theta_0^*, \theta_1^*, \sigma^*)$ for all t; that is, the model is $y_t = \theta_0^* + \theta_1^* y_{t-h} + \epsilon_t;\ \epsilon_t \sim N(0, \sigma^{*2})$. The initial value for the first interval was set to $y_0 = \frac{\theta_0^*}{(1-\theta_1^*)}$. Then, for other intervals, the last value of the previous interval was treated as the initial value of the next interval. This way makes the observations continuous.

- We simulated from the AR(1) process with suddenly and gradually changing parameters in order to investigate the performance of the quantile regression under different types of changes.

- Interval set: $\{I^k\}_{k=1}^5$ for every τ with the following interval lengths,

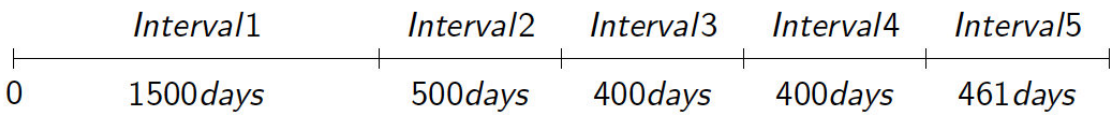


Figure 4.1: Structure of the five different intervals.

Table 4.1: Changes of parameters according to the intervals.

Parameters	Interval 1	Interval 2	Interval 3	Interval 4	Interval 5
θ_0^*	-0.1156	1.1557	-0.1156	0.3467	-0.1156
θ_1^*	0.7827	-0.7827	0.7827	0.6261	0.7827
σ^*	0.5525	0.1000	0.5525	0.4000	0.5525

The true values of the parameters are based on the estimates of an AR(1) model fitted to the $S\&P500$ data from January 2, 1985 to February 4, 2005 [4]. In this scenario, all the parameters are changed according to the time.

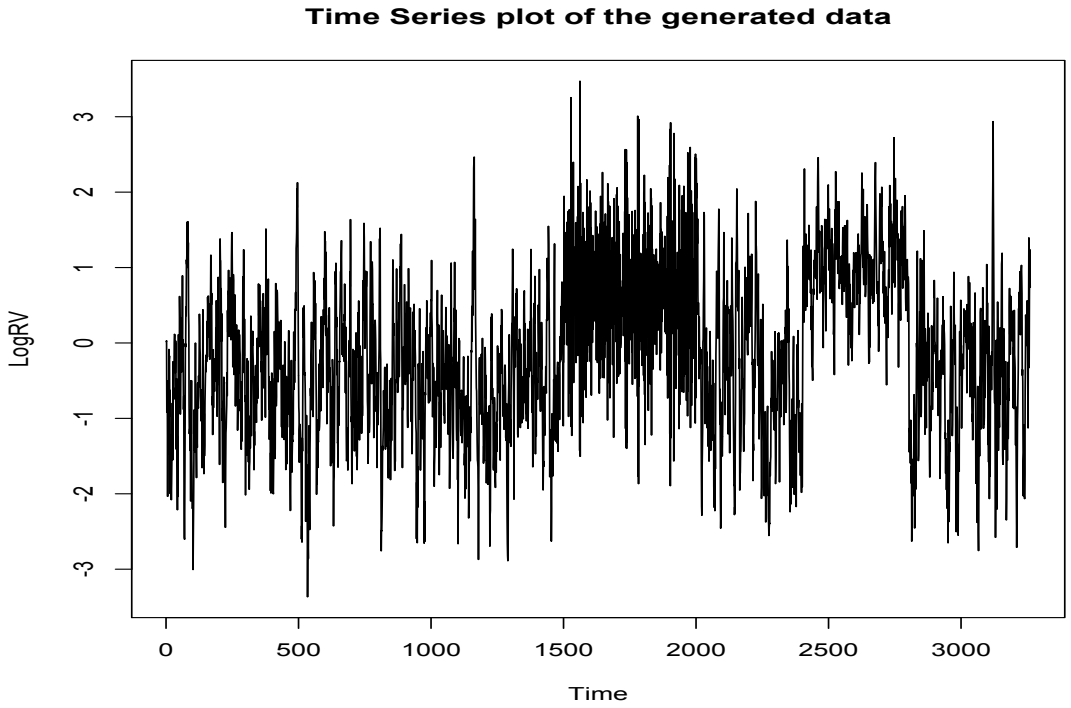


Figure 4.2: Time Series plot of the generated data

Figure 4.2 shows the time series plot of the simulated data. In some time intervals, data follow an upward trend and in some intervals, data do not follow a trend.

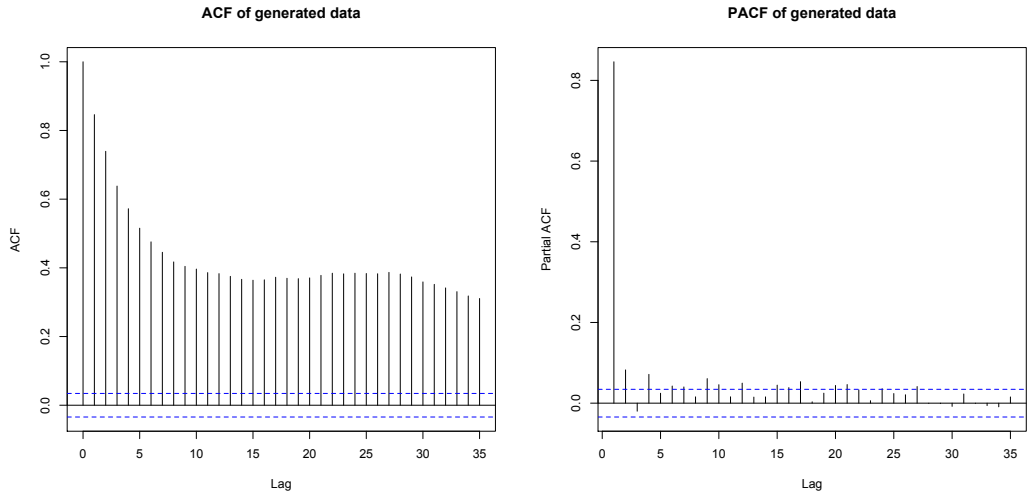


Figure 4.3: ACF and PACF functions for generated data

The autocorrelation function presented in Figure 4.3 has a hyperbolically decaying shape. This is known as a long memory process.

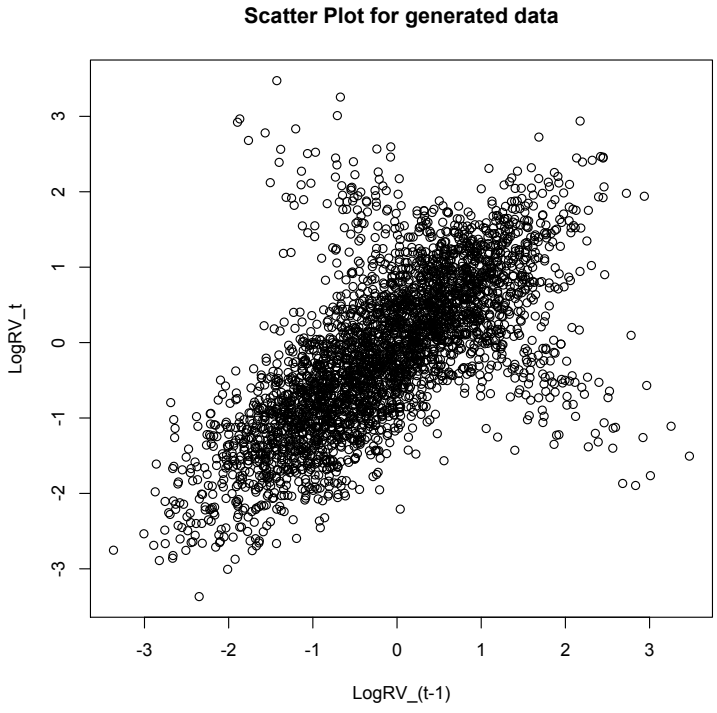


Figure 4.4: Scatter plot for generated data

4.2 Quantile Regression Approach

First, we applied the quantile regression for every interval with different τ values where $(0 < \tau < 1)$. For simplicity, we selected a few important τ values such as

$$\tau = (0.25, 0.5, 0.75, 0.9, 0.95)$$

The quantile regression model is

$$log\ RV_t = \beta_{0t} + \sum_{i=1}^{p} \beta_{\tau t} log\ RV_{t-i} + \epsilon_t,$$

which was introduced in Section 3.3.

We applied the quantile regression method for each interval, and present the results interval-wise for illustration purposes. For each interval, we present a time series plot to describe the shape of the data and a quantile plot that shows the quantile regression lines for selected τ values. Two additional plots are also provided to describe how estimated intercept terms change according to the quantile level (τ) and how estimated slopes change with the quantile level (τ). We generated data using the model $y_t = \theta_0 + \theta_1 y_{t-1} + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. When we fitted the quantile model theoretically, we assumed that the τ^{th} quantile of ϵ is zero. Practically it is not zero. Thus, we corrected that error by adding the τ^{th} percentile of a normal distribution with mean 0 and standard deviation σ.

4.2.1 Quantile Regression Model for Interval I

The Interval-I contains data for 1500 days generated from the model $Y_t = -0.1156 + 0.7827 * Y_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, 0.5525^2)$ and $t = 1, 2, \ldots, 1500$. Figure 4.5 shows time series plot for interval-I data. As expected, there is no upward or downward trend present in the data. Next, we present quantile regression plot for Interval-I data for five different τ values.

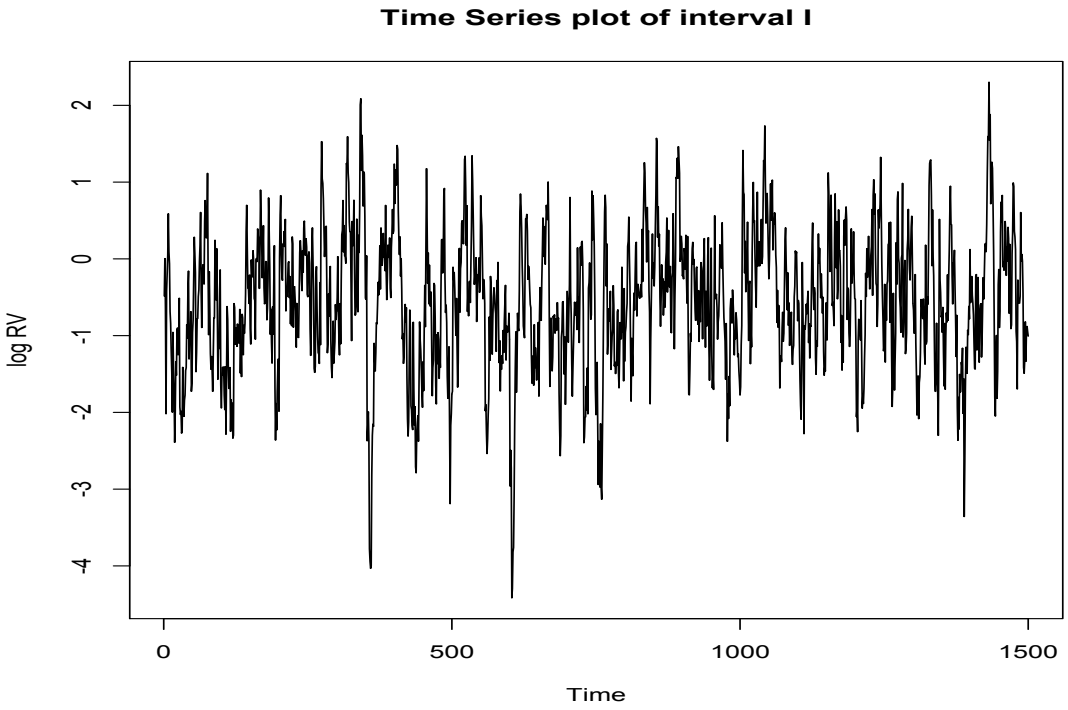


Figure 4.5: Time Series plot for interval-I data

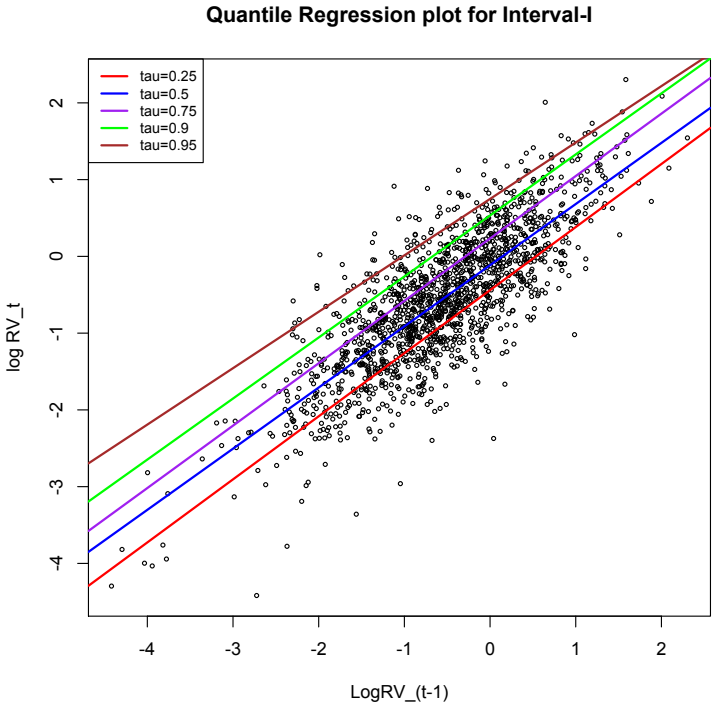


Figure 4.6: Quantile plot of interval-I data

Figure 4.6 shows the quantile plot of the data for Interval-I. All the quantile regression lines are approximately parallel to each other.

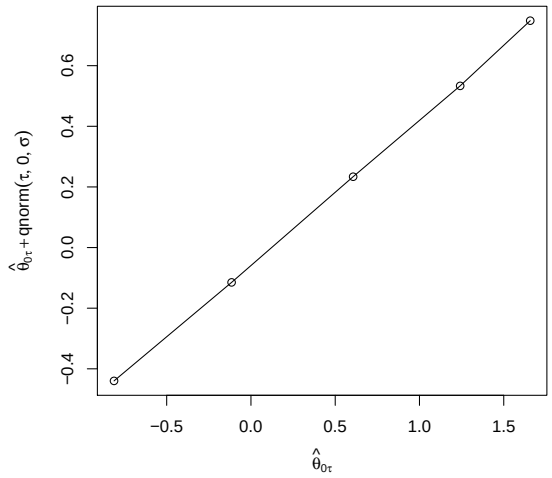


Figure 4.7: Intercept change with τ

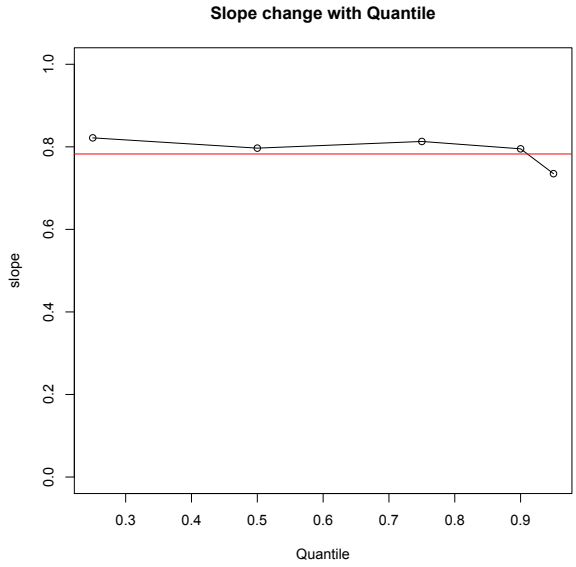


Figure 4.8: Slope change with τ; red line is the true parameter value

Figures 4.7 and 4.8 present the estimated intercept terms and changes in the estimated slopes with respect to τ. In Figure 4.8, the red line indicates the true value of the

estimated parameters. That implies the parameters estimated using the quantile regression method are also very close to the true value.

4.2.2 Quantile Regression Model for Interval II

This interval contains 500 days with parameter values θ_0=1.1557, θ_1=-0.7827, and σ^*=0.1. The model is, $Y_t = 1.1557 - 0.7827 * Y_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, 0.1^2)$ and $t = 1501, 1502, \ldots, 2000$.

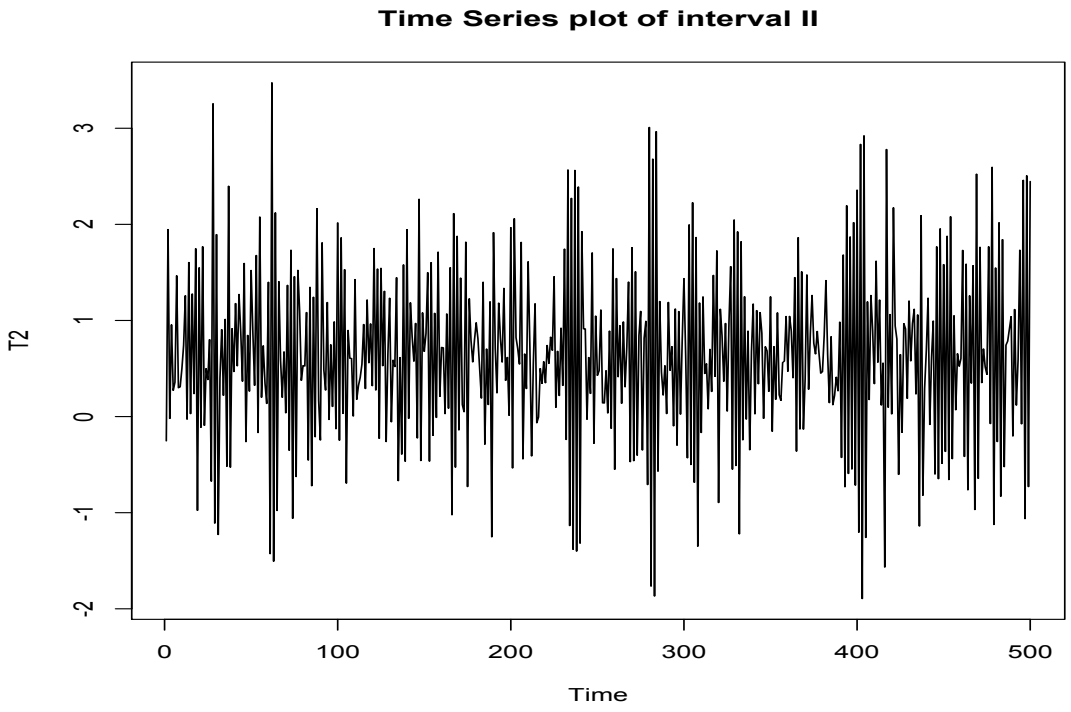


Figure 4.9: Time series plot for interval-II data

Figure 4.9 shows time series plot for interval-II data. There is no upward or downward trend present in the Interval-II data.

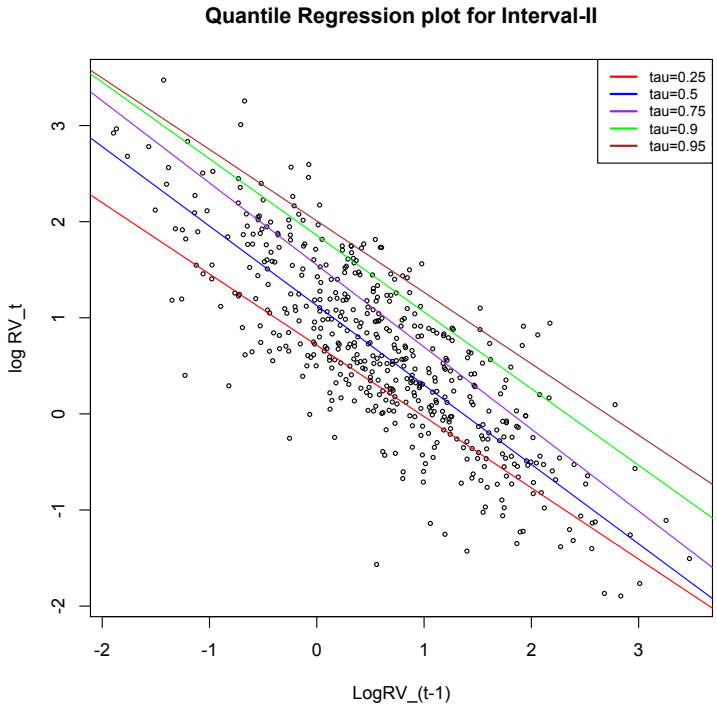


Figure 4.10: Quantile regression plot for interval-II data

Figure 4.10 shows the quantile plot of the data for Interval-II. All the quantile regression lines are approximately parallel to each other.

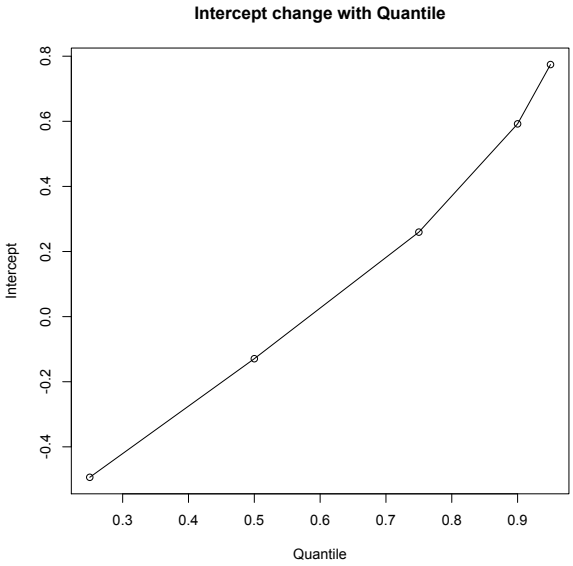


Figure 4.11: Intercept change with τ

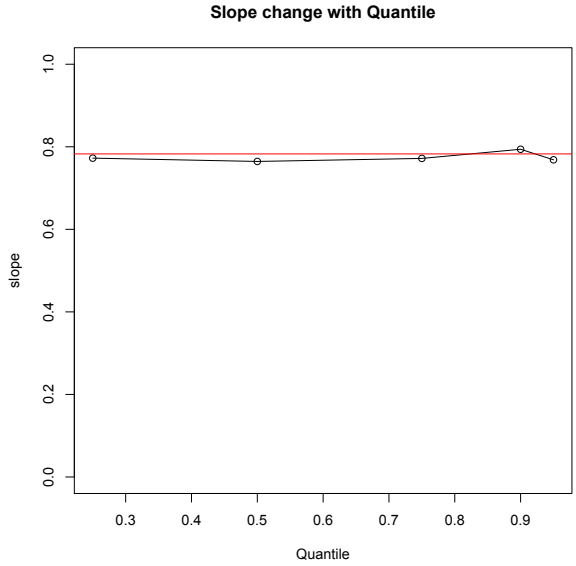


Figure 4.12: Slope change with τ; red line is the true parameter value

Figure 4.11 present the estimated intercept terms change with respect to τ. Figure 4.12 present estimated slopes change with respect to τ. In Figure 4.12, the red line indicates the true value of the estimated parameters. That implies the parameters estimated using the quantile regression method are also very close to the true value.

4.2.3 Quantile Regression for Interval III

This interval contains 400 days with parameter values θ_0=-0.1156, θ_1=0.7827, and σ^*=0.5525. The model is $Y_t = -0.1156 + 0.7827 * Y_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, 0.5525^2)$ and $t = 2001, 2002, \ldots, 2400$. Figure 4.13 shows time series plot for interval-III data. The pattern of the Interval-III data is different from Interval-I and Interval-II. In the first part of the plot, we can see a slight upward trend. In the middle, we can see a downward trend and again slight upward trend in the last part of the plot.

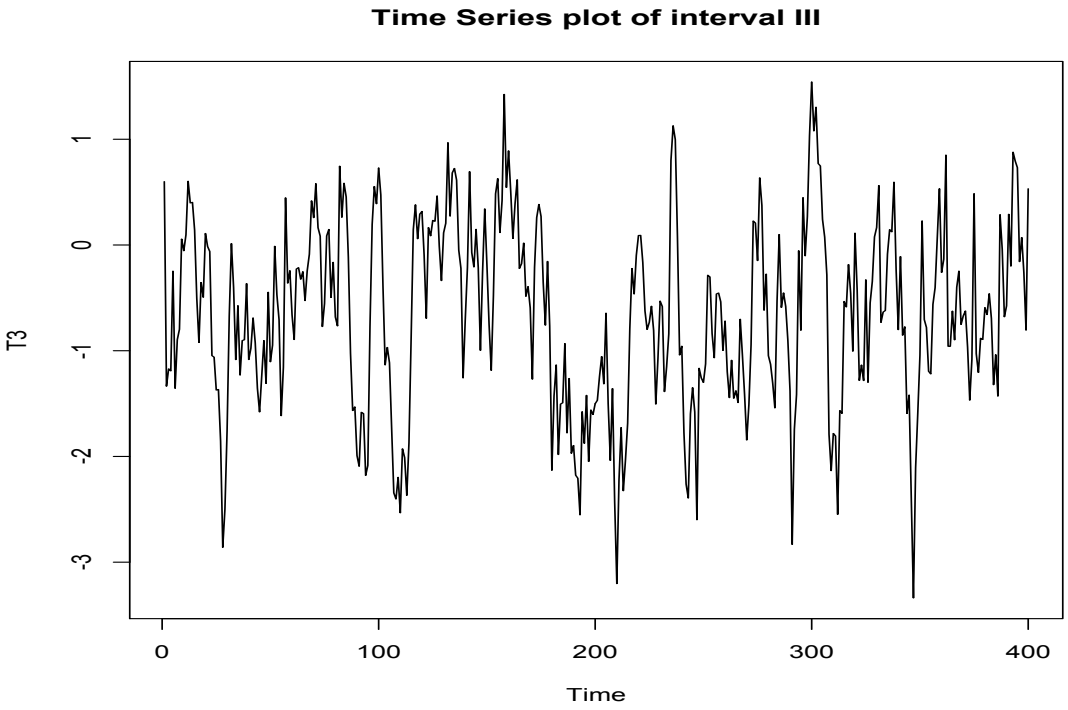


Figure 4.13: Time series plot for interval-III data.

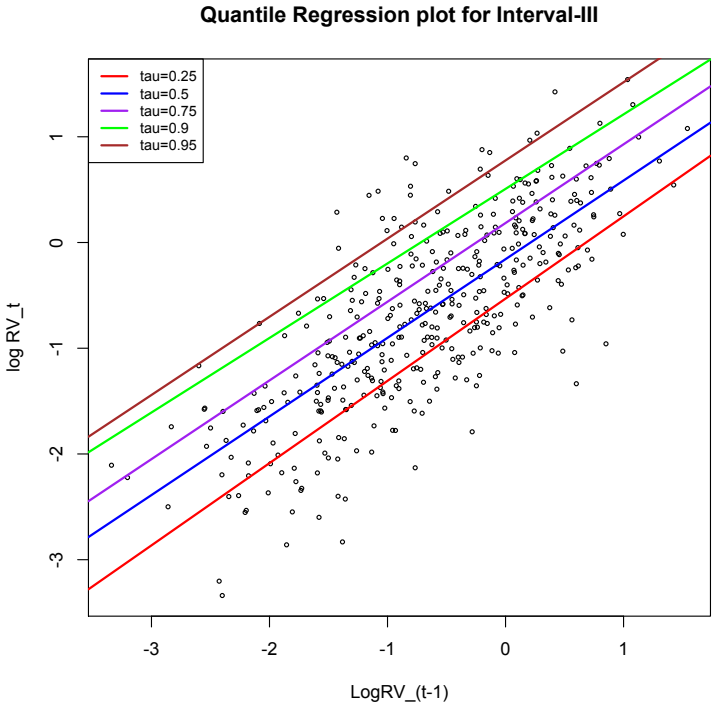


Figure 4.14: Quantile regression plot for interval-III data.

Figure 4.14 shows the quantile plot of the data for Interval-III. All the quantile regression lines are approximately parallel to each other.

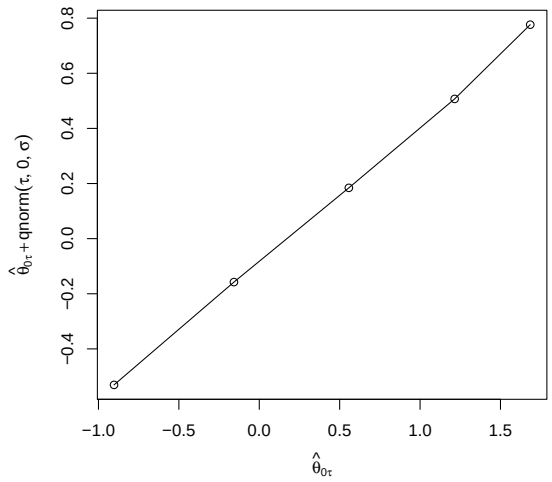


Figure 4.15: Intercept change with τ

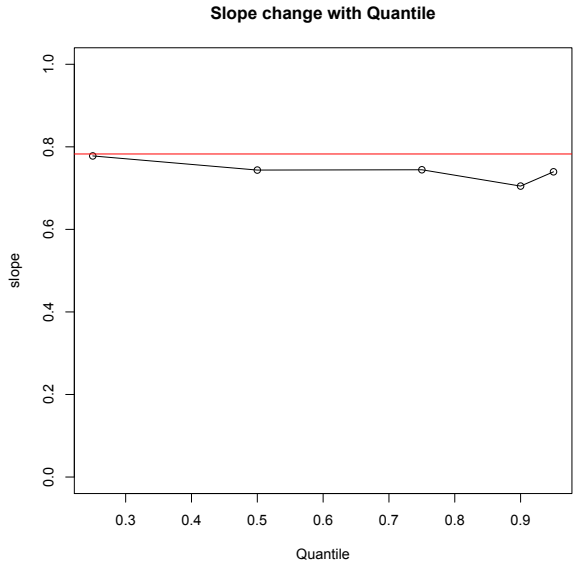


Figure 4.16: Slope change with τ; red line is the true parameter value

Figures 4.15 and 4.16 present the estimated intercept terms and changes in the estimated slopes with respect to τ. In Figure 4.16, the red line indicates the true value of

the estimated parameters. That implies the parameters estimated using the quantile regression method are also very close to the true value.

4.2.4 Quantile Regression for Interval IV

This interval contains 400 days with parameter values θ_0=0.3467, θ_1=0.6261, and σ^*=0.4. The model is $Y_t = 0.3467 + 0.6261 * Y_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, 0.4^2)$ and $t = 2401, 2402, \ldots, 2800$.

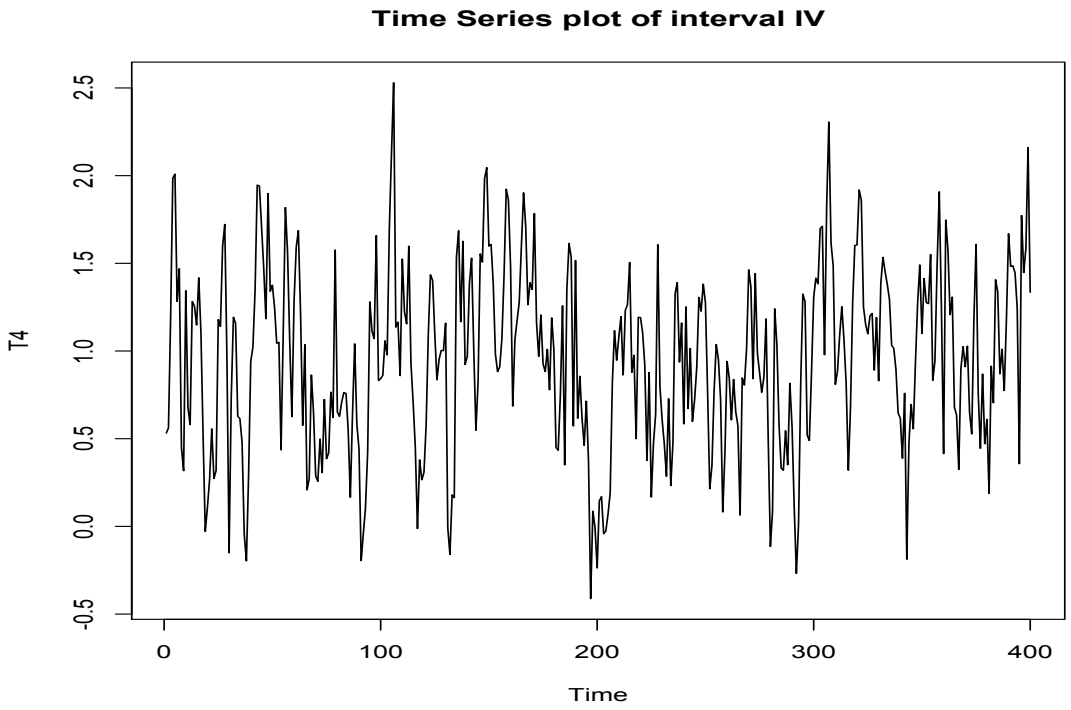


Figure 4.17: Time series plot for interval-IV data.

Figure 4.17 shows time series plot for interval-IV data. There is no upward or downward trend. In the middle part of the plot, we can see a slight fluctuation from normal pattern.

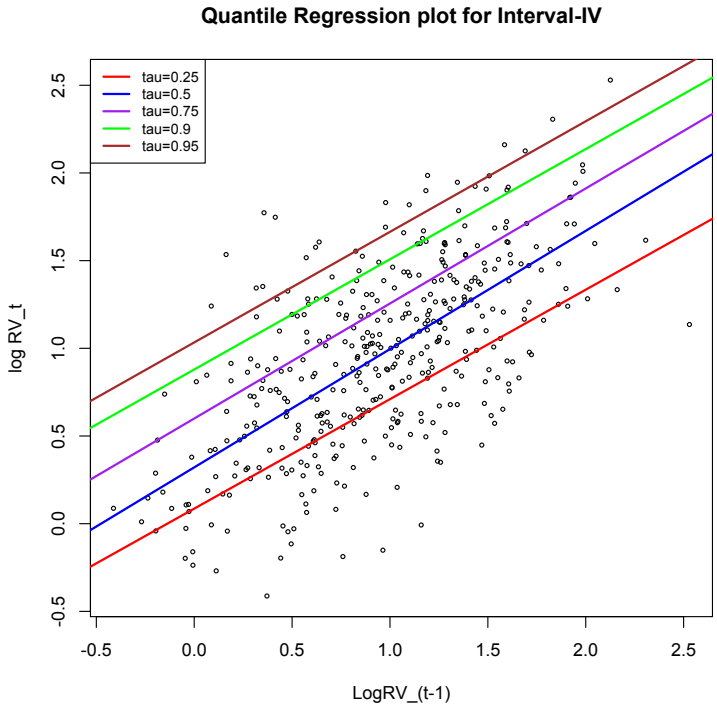


Figure 4.18: Quantile plot for interval-IV data.

Figure 4.18 shows the quantile plot of the data for Interval-IV. All the quantile regression lines are approximately parallel to each other.

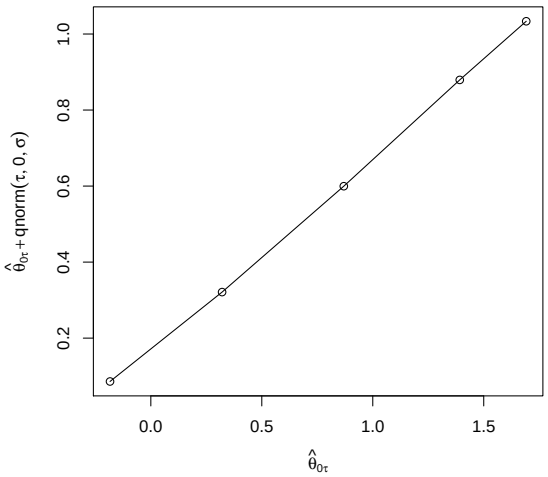


Figure 4.19: Intercept change with τ

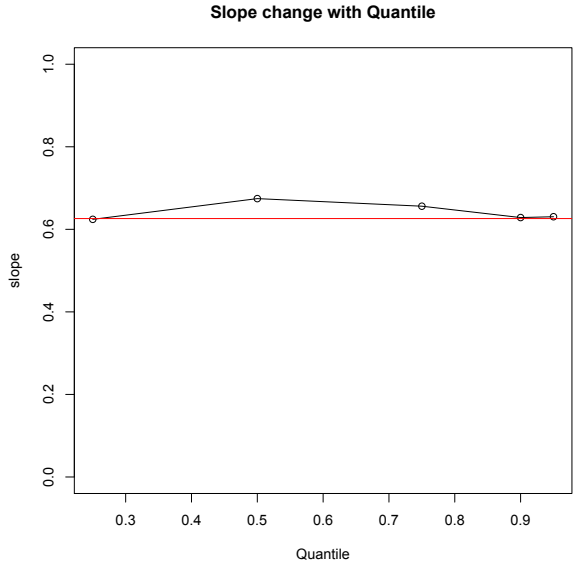


Figure 4.20: Slope change with τ; red line is the true parameter value

Figures 4.19 and 4.20 present the estimated intercept terms and changes in the estimated slopes with respect to τ. In Figure 4.20, the red line indicates the true value of the estimated parameters. That implies the parameters estimated using the quantile regression method are also very close to the true value.

4.2.5 Quantile Regression for Interval V

This interval contains 461 days with parameter values θ_0=-0.156, θ_1=0.7827, and σ^*=0.5525. The model is, $Y_t = -0.156 + 0.7827 * Y_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0, 0.5525^2)$ and $t = 2801, 2802, \ldots, 3261$. Figure 4.21 shows time series plot for Interval-V data and we can observe a slight downward trend here. .

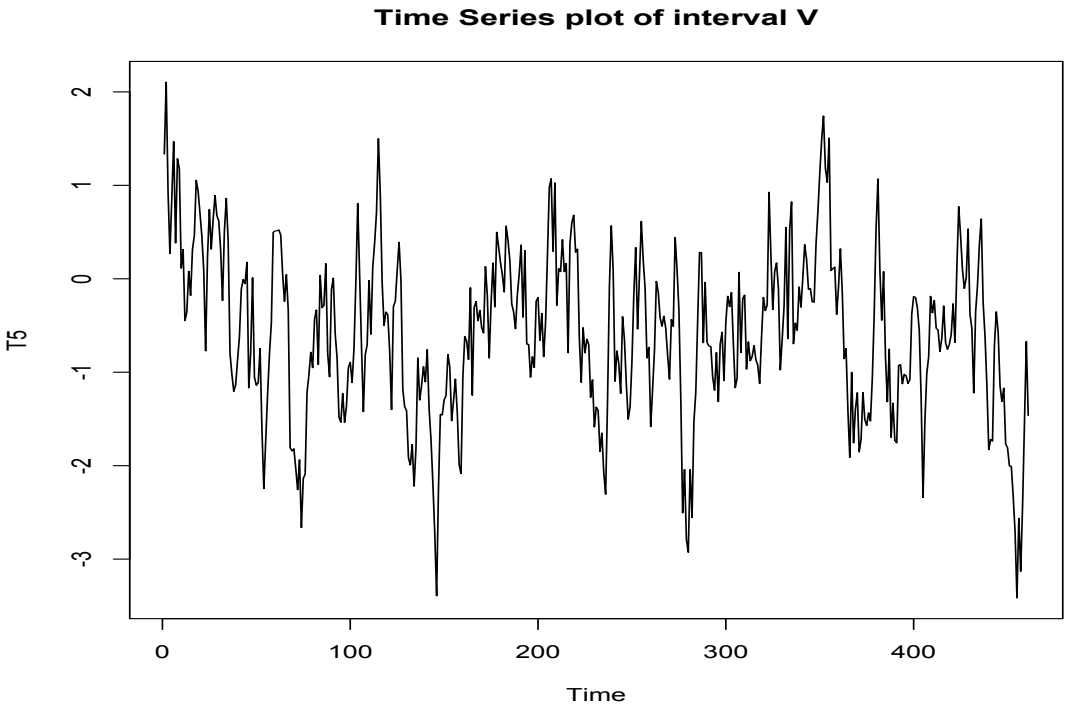


Figure 4.21: Time series plot for interval-V

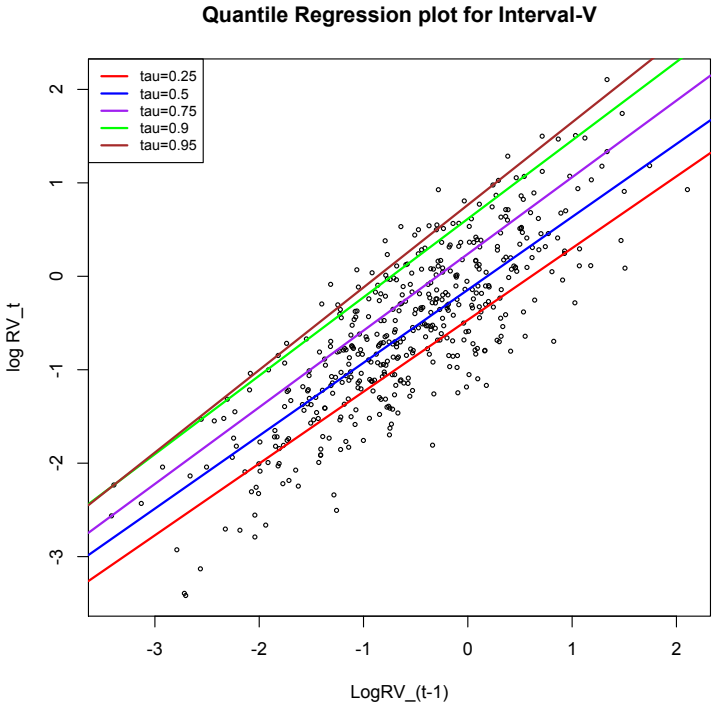


Figure 4.22: Quantile plot for interval-V data

Figure 4.22 shows the quantile plot of the data for Interval-V. All the quantile regression lines are approximately parallel to each other.

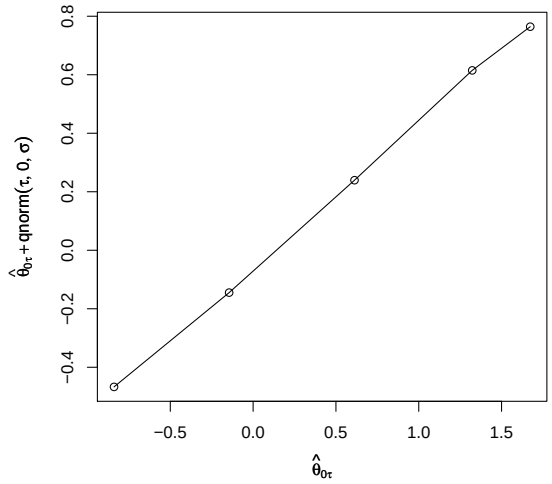


Figure 4.23: Intercept change with τ

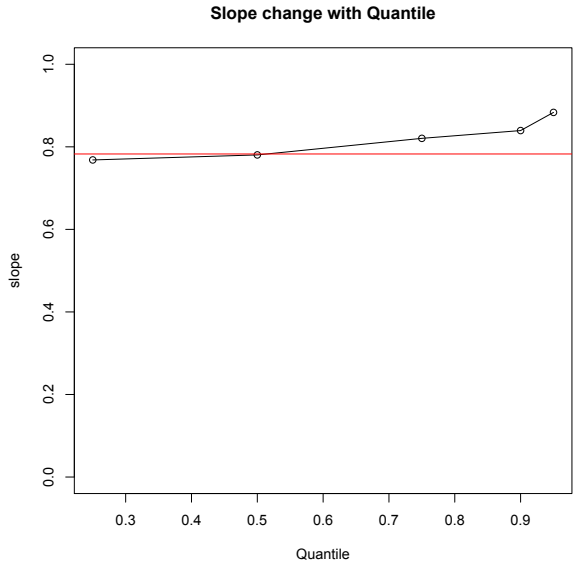


Figure 4.24: Slope change with τ; red line is the true parameter value

Figures 4.23 and 4.24 present the estimated intercept terms and changes in the estimated slopes with respect to τ. In Figure 4.24, the red line indicates the true value of

the estimated parameters. That implies the parameters estimated using the quantile regression method are also very close to the true value.

The first part of the simulation study was carried out to find the performance of quantile regression on data with different characteristics. We generated data for five intervals, where three of them (Interval-I, Interval-II, and Interval-V) have the same set of parameters but different lengths, while the other two intervals have totally different parameter (Interval-II and Interval-IV) sets and different lengths. This way we can investigate the performance of the quantile regression under different types of changes. All the lines in the quantile plots are approximately parallel to each other and quantile regression estimators are very close to true values. Based on these facts and by looking at all the graphs, we can say that the quantile regression method describes the generated data well for all intervals. As the regression model differs from interval to interval, it is inappropriate to predict the volatility using one general model. It will be critical to identify the interval of homogeneity for accurate predictions.

4.3 Identification of the Interval of Homogeneity

Volatility clusters through time. Therefore, within some time periods, volatility tends to have a similar pattern. These time periods are called intervals of homogeneity. This section presents another simulation study which is carried out to find the longest interval of homogeneity. The purpose of this simulation study is to illustrate how to select the longest interval of homogeneity, which makes the quantile regression model a good approximation to the data. To determine the interval of homogeneity, we use likelihood ratio testing techniques.

4.3.1 Simulation Setup

In this simulation study, we focus on the first two intervals of the above simulation study. Interval-I contains 1500 data and interval-II contains 500 data. In the previous simulation study, we used different parameters in each interval to generate heterogeneous data sets. The study is carried out as follows,

1. Initially select a τ (quantile level) value.

2. Let interval-I data be y_t, and then fit a quantile regression model, then find quantile regression parameters $\hat{\theta_1}$ at τ^{th} quantile level. Then, the quantile regression is $y_t = \theta_{01} + \theta_{11} y_{t-1} + \epsilon_1$, where $t = 1, 2, \ldots, 1500$. For simplicity we let $\underset{\sim}{y_t} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{1500} \end{pmatrix}$, and $\underset{\sim}{\theta_1} = (\theta_{01}, \theta_{11})^T$

3. For interval-II, the estimated quantile regression parameter set is $\hat{\theta_2}$ at the τ^{th} quantile level. We took $y_{1501} = y_{1500}$ because this way we can make the observations continuous. Then, the quantile regression model for interval-II is, $y_t = \theta_{02} + \theta_{12} y_{t-1} + \epsilon_2$, where $t = 1501, 1502, \ldots, 2000$. Similary, we let $\underset{\sim}{y_t} = \begin{pmatrix} y_{1501} \\ y_{1502} \\ \vdots \\ y_{2000} \end{pmatrix}$, and $\underset{\sim}{\theta_2} = (\theta_{02}, \theta_{12})^T$

4. We combined interval-I and interval-II to form the full interval, and the quantile regression parameter set is θ. Then, the quantile regression model for the full interval is $y_t = \theta_{00} + \theta_{10} y_{t-1} + \epsilon_0$, where $t = 1, 2, \ldots, 2000$, and we let $\underset{\sim}{y_t} = \begin{pmatrix} y_0 \\ y_2 \\ \vdots \\ y_{1500} \\ y_{1500} \\ \vdots \\ y_{2000} \end{pmatrix}$, and $\underset{\sim}{\theta_0} = (\theta_{00}, \theta_{10})^T$

5. Next, calculate the likelihood ratio (LR) statistics based on asymmetric Laplace distribution. Take this likelihood ratio statistic as LR_1. This method can be explain as follows; Suppose the $y_i \sim ALD(x_i^T \theta_\tau, \sigma; \tau),\ i = 1, 2, \ldots, n$, are independent. Then, the likelihood function for n observations is

$$L(\theta, \sigma | y) \propto \frac{1}{\sigma^n} exp\Big\{ -\sum_{i=1}^{n} \rho_\tau \Big(\frac{y_i - X_i^T \theta_\tau}{\sigma} \Big) \Big\}, \tag{4.1}$$

where $X_i = (1, y_i)$, σ can be estimated using $\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(y_i - X_i^T\hat{\theta}_\tau\big)$ and $\propto$ means $\mathrm{L}(\cdot)$ is proportional. Let T_1 denote the time interval-1 and n_1 is the length of the time interval-1. Likewise, T_2 denote the time interval-II and length of time interval-II is n_2. Then LR values for interval-I, interval-II, and full interval can be given as,

$$L_{\theta_1}(T_1) \propto \sigma_1^{-n_1} exp\{-n_1\}, \tag{4.2}$$

$$L_{\theta_2}(T_2) \propto \sigma_2^{-n_2} exp\{-n_2\}, \tag{4.3}$$

$$L_{\theta_0}(T_1, T_2) \propto \sigma_0^{-N} exp\{-N\}, \tag{4.4}$$

where $N = n_1 + n_2$. Hypothesis of this test is

$$H_0 : \underset{\sim}{\theta_1} = \underset{\sim}{\theta_2}, H_a : \underset{\sim}{\theta_1} \neq \underset{\sim}{\theta_2} \tag{4.5}$$

The log likelihood ratio statistic for testing the hypothesis (4.5) is given by,

$$\begin{aligned}
LR_1 &= -2\log\left(\frac{max_{\{\theta_1,\sigma\}\in H_0}\{L_{H_0}(T_1,T_2)\}}{max_{\{\theta_1,\theta_2,\sigma\}\in H_a}\{L_{H_a}(T_1), L_{\theta_2}(T_2)\}}\right) \\
&= -2\log\left(\frac{\hat{\sigma_0}^{-N}\exp\{-N\}}{\hat{\sigma_1}^{-n_1}\exp\{-n_1\}\hat{\sigma_2}^{-n2}\exp\{-n_2\}}\right) \\
&= -2\big(\log(\hat{\sigma_0}^{-N}) - \log(\hat{\sigma_1}^{-n_1}) - \log(\hat{\sigma_2}^{-n_2}) - N + n_1 + n_2\big) \\
&= -2\big(\log(\hat{\sigma_0}^{-N}) - \log(\hat{\sigma_1}^{-n_1}) - \log(\hat{\sigma_2}^{-n_2})\big), \quad where\ N = n_1 + n_2 \\
&= 2\big(\log(\hat{\sigma_0}^{N}) - \log(\hat{\sigma_1}^{n_1}) - \log(\hat{\sigma_2}^{n_2})\big) \\
&= 2\log\big(\frac{\hat{\sigma}^N}{\hat{\sigma_1}^{n_1}\hat{\sigma_2}^{n_2}}\big).
\end{aligned}$$

6. Then, find the empirical distribution of the likelihood ratio statistic using re-sampling techniques.

   - Fit a linear regression model for y_t and y_{t-1} then find ordinary least squares (OLS) estimates of the regression parameters (Let $\hat{\beta}_0, \hat{\beta}_1$ be OLS estimates of the parameters) for interval-I (for 1500 data). Then the OLS prediction equation is,
   $$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 y_{t-1},\ t = 1, 2, \ldots, 1500. \tag{4.6}$$
   - Use parameters obtained from the previous model (equation (4.25)) to

predict the next 500 data. Take $\tilde{y}_t$ as predicted values,

$$\tilde{y}_t = \hat{\beta}_0 + \hat{\beta}_1 y_{t-1}, \quad t = 1501, 1502, \ldots, 2000. \tag{4.7}$$

- Find residuals from the equation (4.6) and fit another linear regression model for interval-II and find the residuals. Combine all residuals, and then we have 2000 residuals. Let e_t be the residuals of linear regression models. The first 1500 residuals are based on equation (4.6) and the last 500 residuals are coming from the linear regression model which we fit for interval-II.
- Combine fitted values, i.e. $\hat{y}_i$ and $\tilde{y}_i$.

  $\hat{y}_i,\ i = 1, 2, \ldots, 1500.$

  $\tilde{y}_i,\ i = 1501, 1502, \ldots, 2000.$
- Permute $\{e_i\}$. Let $\tilde{e}_i$ be the permuted residuals, $i = 1, 2, \ldots, 2000$.
- Create new data by adding fitted values and permuted residuals. Let new y values be Y^*,

$$Y^* = \begin{pmatrix} \hat{y}_1 + \tilde{e}_1 \\ \hat{y}_2 + \tilde{e}_2 \\ \vdots \\ \hat{y}_{1500} + \tilde{e}_{1500} \\ \tilde{y}_{1501} + \tilde{e}_{1501} \\ \vdots \\ \tilde{y}_{2000} + \tilde{e}_{2000} \end{pmatrix},$$

- Break the new Y^* into two parts such that the first 1500 as Y_1^* and the last 500 as Y_2^*.
- Then, we carry out the likelihood ratio test based on ALD for new data that is $Y^*,\ Y_1^*$, and Y_2^*.
- Repeat these steps 1000 times and find 1000 LR values. That is the empirical distribution of LR values.
- Take $100 \times (1-\alpha)^{th}$ percentile of 1000 LR values as the ζ_k.
- Compare the ζ_k and LR_1 to test the hypothesis. The null hypothesis is rejected if the LR_1 value is greater than the ζ_k value.

4.4 Results

In this section, we present the results of the simulation study. First, we obtained results for different α values. Table 4.2 shows the results for selected α values.

Table 4.2: Likelihood ratio test results for different α values.

α	LR_1	ζ	Decision
0.05	0.640	0.214	Reject H_0
0.01	0.640	0.222	Reject H_0

First, we selected $\alpha = 0.05$. Then ζ is the 95% percentile of 1000 likelihood ratio values. The LR_1 value is greater than the ζ value, which implies that we reject the null hypothesis. As the second step, we selected α value as 0.01, and the ζ value is the 99% percentile of the 1000 likelihood ratio values. In this case, we obtained the same conclusion as the first case. Next, we selected a few different quantiles for the test. The results are shown below,

Table 4.3: Likelihood ratio test results for different τ values.

τ	LR_1	ζ	Decision
5%	0.616	0.202	Reject H_0
95%	0.653	0.150	Reject H_0
50%	0.640	0.213	Reject H_0
85%	0.614	0.156	Reject H_0

The null hypothesis ($H_0 : \underset{\sim}{\theta_1} = \underset{\sim}{\theta_2}, \ \ \sigma > 0$) was rejected for all selected τ values. Interval-I data were generated using $Y_t = a_1 + b_1 Y_{t-1} + \epsilon_1$, where $\epsilon_1 \sim N(0, \sigma_1^2)$ and $t = 1, 2, \ldots, 1500$. Interval-II data were generated using $Y_t = a_2 + b_2 Y_{t-1} + \epsilon_2$, where $\epsilon_1 \sim N(0, \sigma_2^2)$ and $t = 1501, 1502, \ldots, 2000$. For this part of the study, we selected the same values for a_1 and a_2. Also, we chose one value for both standard deviations of ϵ_1 and ϵ_2. That is, we generated data with the same characteristics; we fixed a and changed b_i in the model Y_t=a+$b_i Y_{t-1}$+ϵ_i where $\epsilon_i \sim$ N(0,σ_i^2) and i=1,2

Initially we have two different data sets, but when b_1 was close to b_2, all the parameters were the same. If our simulation set up is correct, the p-values should be

greater than 0.05 when b_1 and b_2 are getting close to each value. When b_1 is very close to b_2, we can use one model to describe the data in both intervals. According to the results, it is clear that our simulation setup can be used to identify homogeneous time intervals.

Table 4.4: Fixed a and same variances and different b_1, b_2.

	Value
a	-0.1156
b_1	0.7827
b_2	-0.7827
σ_1	0.1
σ_2	0.1
LR_1	0.633
ζ	0.370
$P-value$	0.000

Table 4.5: Fixed a and same variances and different b_1, b_2.

	Value
a	1.1557
b_1	0.7827
b_2	-0.7827
σ_1	0.5525
σ_2	0.5525
LR_1	0.857
ζ	0.567
$P-value$	0.000

Table 4.6: Fixed a and same variances and different b_1, b_2.

	Value
a	1.1557
b_1	0.7827
b_2	-0.7827
σ_1	0.1
σ_2	0.1
LR_1	1.040
ζ	0.765
$P-value$	0.000

Table 4.7: Fixed a and same variances and b_1 and b_2 are close.

	Value
a	-0.1156
b_1	0.7827
b_2	0.500
σ_1	0.1
σ_2	0.1
LR_1	0.048
ζ	0.117
$P-value$	0.050

Table 4.8: Same parameters for both intervals.

	Value
a	-0.1156
b_1	0.7827
b_2	0.7827
σ_1	0.1
σ_2	0.1
LR_1	0.067
ζ	0.128
$P - value$	0.546

According to the estimated p-values of all the tests, we can conclude that, when two intervals have data with different characteristics, the test rejects the null hypothesis. This result shows that that is we need two different models to capture the behavior of the data. In order to cross-validate our simulation study, we generated data with same characteristics by selecting two close values for b_1 and b_2. According to the results presented in Tables 4.7 and 4.8 we fail to reject the null hypothesis that supports our claim that one model can be used to explain the data. This implies that our simulation setup works well to find the longest interval of homogeneity.

Next, we selected 17 different values for b_2 while keeping the b_1 value as zero to estimate the power function of the hypothesis test. Also we selected the same value for both a_1 and a_2, as well as σ_1 and σ_2. Here we assume that the alternative hypothesis is true. Then, we generated the data for interval-I and interval-II with to the parameters as given in table 4.9. We recorded whether the test rejected the null hypothesis or not. This procedure carried out 500 times and the power of the test is calculated as,

$$power = \text{Number of rejects}/500. \tag{4.8}$$

Results are given in the following table for selected $b_1 - b_2$ values,

Table 4.9: Values of the power for different $b_1 - b_2$ values.

b_1	b_2	$\delta = b_1 - b_2$	power of the test
	-0.19654	0.19654	0.974
	-0.16454	0.16454	0.930
	-0.15654	0.15654	0.912
	-0.140886	0.140886	0.822
	-0.125232	0.125232	0.732
	-0.109578	0.109578	0.590
	-0.093924	0.093924	0.422
0.0000	-0.07827	0.07827	0.342
	-0.070443	0.070443	0.298
	-0.062616	0.062616	0.208
	-0.054789	0.054789	0.186
	-0.046962	0.046962	0.162
	-0.039135	0.039135	0.118
	-0.031308	0.031308	0.092
	-0.015654	0.015654	0.070
	-0.007827	0.007827	0.068
	0.00000	0.00000	0.064

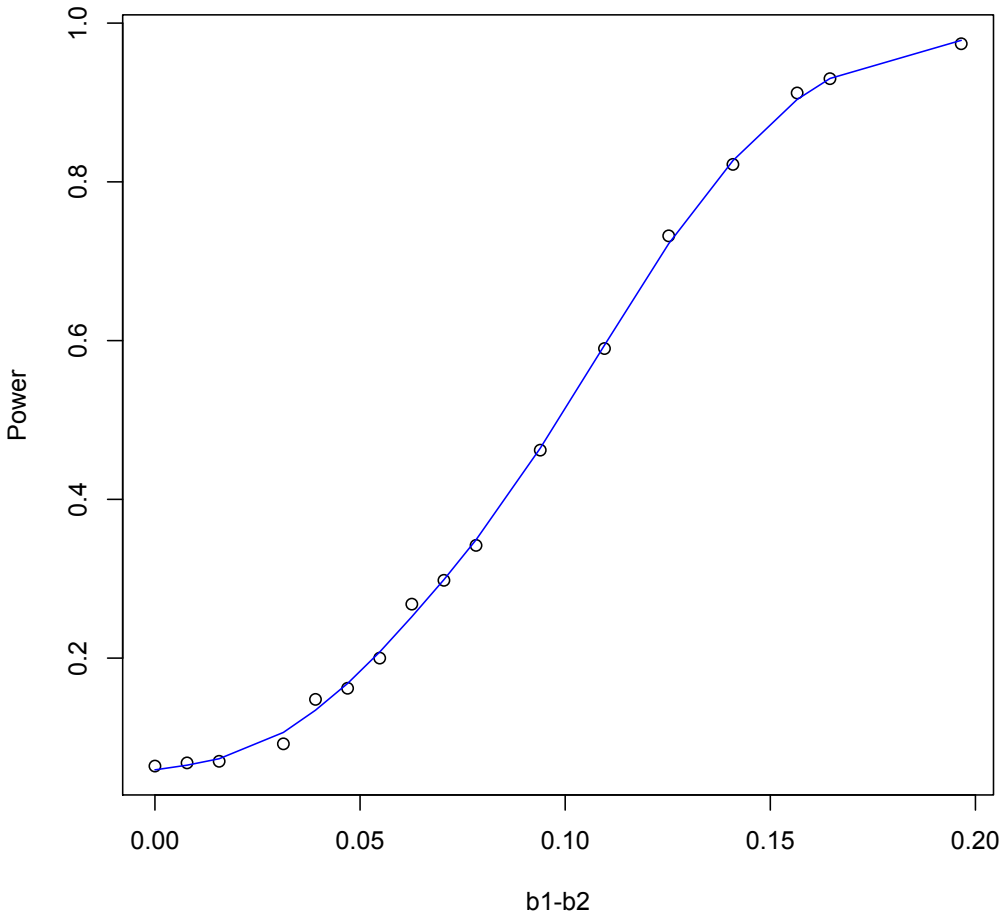


Figure 4.25: Power function of the test.

In the Figure 4.25, x-axis shows the difference of b_1 and b_2. When the difference of b_1 and b_2 are small, power is also small.

4.4.1 Asymptotic Distribution of the Estimators of the Quantile Regression Parameters

Next, we ran a few tests to check whether the estimators of the quantile regression parameters are asymptotically normally distributed. First, we perform the Shapiro-Wilk test. The p-value of the Shapiro-Wilk test is 0.6256, which indicates that we fail to reject the normality assumption.

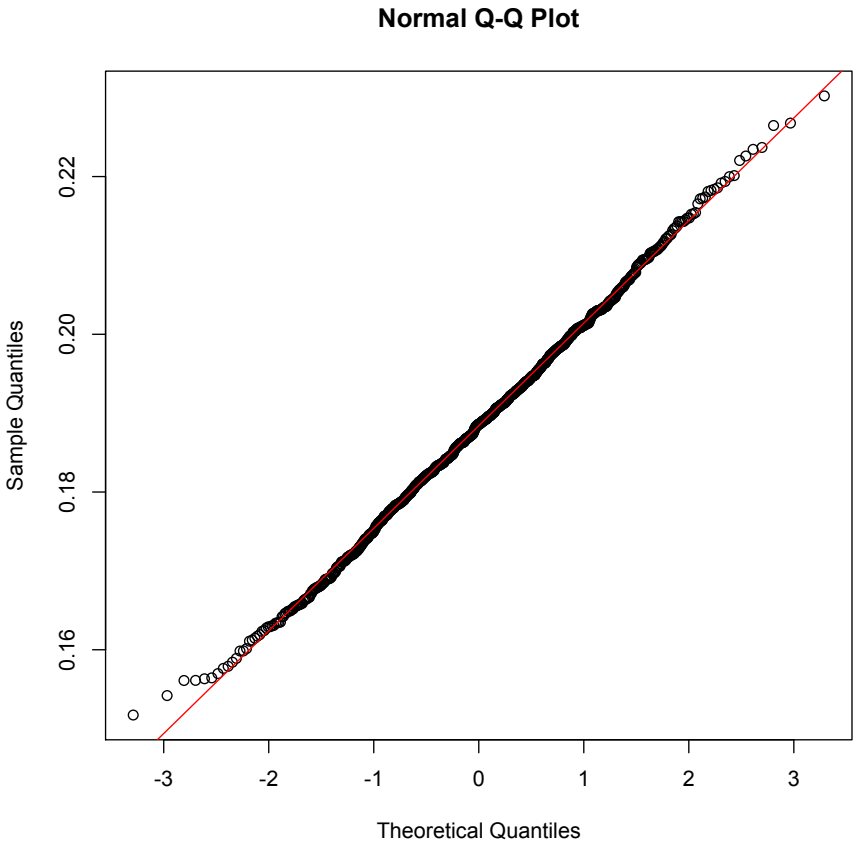


Figure 4.26: qqplot for estimates of quantile regression parameters.

Also, qqplot depicted in Figure 4.26 that the normal disribution is adequate for data. Taking all these factors into consideration, we can conclude that the estimators of the quantile regression parameters are asymptotically normally distributed.

Chapter 5

Summary and Future Work

5.1 Summary

In this thesis, we introduce a quantile regression model for the volatility of assets. This study was inspired by the article "Localized Realized Volatility Modeling" by Chen et al. [4]. They proposed a new model called the Localized Autoregressive (LAR) model to forecast financial volatility. Identifying the interval of homogeneity is the first step of their approach. Volatility is not constant, and clusters through time. Within some time periods, volatility tends to have a similar pattern. These time periods are called intervals of homogeneity. Our approach attempts to identify the longest interval of homogeneity using past data. Then, we apply the quantile regression model separately for each interval.

In the first part of our simulation study in Chapter 4, we focus on generating data with some structural changes. To illustrate the quantile regression approach, we draw quantile regression plots for each interval. Most of the data lie between the 0.25^{th} and 0.95^{th} quantile levels (τ) of $\log RV_t$. This shows us that the proposed quantile regression model is appropriate to describe the data. For this method, we do not need any distributional assumptions. As a result, we can directly interpret the results at selected quantiles. This might be more interesting to researchers and practitioners in the area of finance.

In the second part of the study, we illustrated how to identify the interval of homogeneity. For that, we used two different datasets which we generated in the first part of the simulation study. This is important because with a real-life data set, identifying these intervals usually a challenge. To find out the length of the interval

of homogeneity, we can use the same steps we carried out in the second part of the simulation study presented in Chapter 4. As the initial step, we should define n_1 and n_2, where n_1 is the length of the first interval and n_2 is the length of the second interval. For example, to start the test, we can take the first five days as the first interval and the rest of the data set as the second interval. If the test does not reject the null hypothesis in equation (4.5), given in Chapter 4, we merge the next data interval with the first interval. Then, we can apply the test again with the merged data set. Likewise, we can do this until we reject the null hypothesis and find the longest interval of homogeneity. After identifying all the homogeneous intervals successfully, we can separately apply the quantile regression method for each interval.

5.2 Challenges and Future Work

In the study, "Localized Realized Volatility Modeling" by Chen et al. [4], the researchers applied the LAR procedure to the S&P 500 futures indices. For all their calculations, they used minute-by-minute data of S&P 500 index futures from January 2, 1985 to February 4, 2005. Finding intraday data is the first challenge we faced in our study. The reason for this challenge is the limited access to minute-by-minute financial data. Most of the institutions allow us to access weekly, monthly, and annual data. There is a cost involved with minute-by-minute data acquisition. Even though we have access to historical intraday data (minute-by-minute data), handling (refining and preprocessing) will be another challenge. Also, finding the interval of homogeneity is a sequential testing procedure, which brings larger computational cost to the analysis.

If intraday data are available, model parameters can be estimated more accurately because multiple data points for a day will show the correct fluctuation of the daily volatility. Therefore, we propose using minute-by-minute financial data to accurately model and forecast the daily volatility, as a further work. Moreover, in the future we can compare predicted values using different models. First, we can predict future values using one quantile regression model for both interval-I and interval-II. Then, we can compare those values with the values predicted using only interval-II model.

Bibliography

[1] R. T. Baillie, T. Bollerslev, and H. O. Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74:3–30, 1996.

[2] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

[3] J. Y. Campbell, A. W. Lo, A. C. MacKinlay, et al. *The econometrics of financial markets*, volume 2. Princeton University Press, Princeton, NJ, 1997.

[4] Y. Chen, W. K. Hardle, and U. Pigorsch. Localized realized volatility modeling. *Journal of American Statistical Association*, 105:1376–1393, 2010.

[5] R. Chou. Volatility persistence and stock valuations : some empirical evidence using garch. *Journal of Applied Econometrics*, 3:279–294, 1988.

[6] R. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflations. *Econometrica*, 50:987, 1982.

[7] R. F. Engle and A. J. Patton. What good is a volatility model? *Quantitative Finance*, 1:237–245, 2001.

[8] E. F. Fama. The behavior of stock market prices. *The Journal of Business*, 38:34–105, 1965.

[9] M. Geraci and M. Bottai. Likelihood based inference for quantile regression using the asymmetric laplace distribution. *Biostatistics*, 8:140–154, 2013.

[10] L. R. Glosten, R. Jagannathan, and D. E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48:1779–1801, 1993.

[11] S. F. Gray. Modeling the conditional distribution of interest rates as a regime-switching. *Journal of Financial Economics*, 42:27–62, 1996.

[12] Y. Huang, Alex. Volatility forecast by quantile regression. *Applied Economics*, 44:423–433, 2012.

[13] R. Koenker. Quantile regression, volume 38 of econometric society monographs, 2005.

[14] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46:33–55, 1978.

[15] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

[16] B. Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36:394–419, 1963.

[17] D. Marcek and L. Falat. Volatility forecasting in financial risk management with statistical models and arch-rbf neural networks. *Journal of Risk Analysis and Crisis Response*, 4:77, 2014.

[18] J. Marcucci. Forecasting stock market volatility with regime-switching garch models. *Studies in Nonlinear Dynamics and Econometrics*, 9, 2005.

[19] D. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1991.

[20] B. Sanchez, H. Lachos, and V. Labra. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Journal of Statistics*, 81:1565–1578, 2007.

[21] G. W. Schwert. Why does stock market volatility change over time? *The Journal of Business*, 44:1115–1153, 1989.

[22] R. S. Tsay. *Analysis of financial time series*, volume 543. John Wiley & Sons, 2005.

[23] K. Yu, Z. Lu, and J. Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician),*, 52:331–350, 2003.

[24] J. M. Zakoian. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, 18:931–955, 1994.